Handbook for Assessment in the Service of Learning

Volume I

Foundations for Assessment in the Service of Learning

Edited by
Eric M. Tucker
Eleanor Armour-Thomas
Edmund W. Gordon

Handbook for Assessment in the Service of Learning

Volume I

Foundations for Assessment
in the Service of Learning

UMassAmherst
University Libraries

Edited by
Eric M. Tucker
Eleanor Armour-Thomas
Edmund W. Gordon

Series Editors

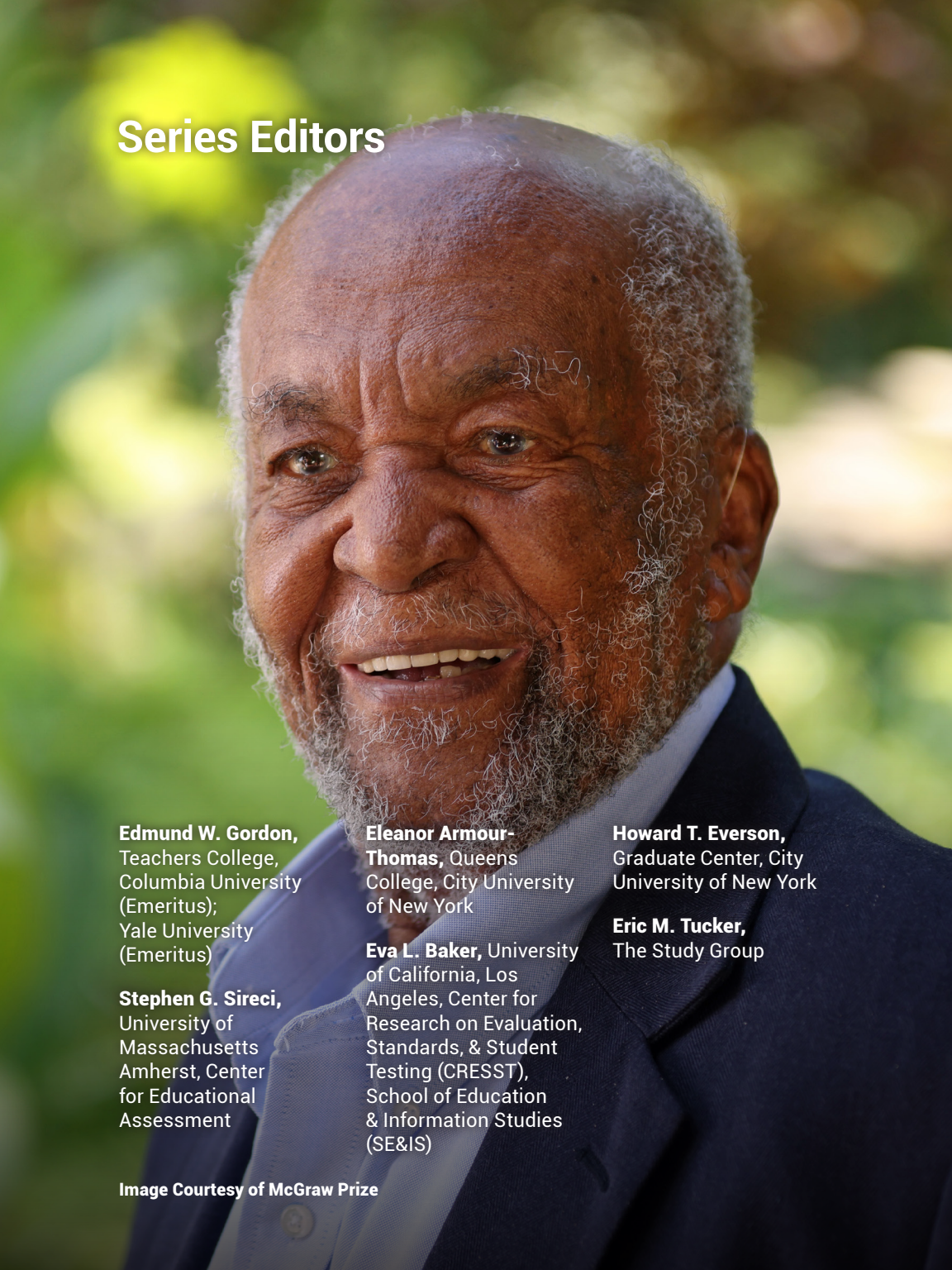


Image Courtesy of McGraw Prize

UMassAmherst

University Libraries

Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning ©

First edition published September 2025
by the University of Massachusetts Amherst Libraries
https://openpublishing.library.umass.edu/

DOI: 10.7275/2h95-jf35
ISBN: 978-1-945764-33-2

Cover Design by Dezudio
Book Design by The Study Group

The suggested citation for this handbook is: Tucker, E. M., Armour-Thomas, E., & Gordon, E. W. (Eds.). (2025). *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Contents

Volume Contributors

Eleanor Armour-Thomas, Queens College, City University of New York (Emeritus)

Aneesha Badrinarayan, Education First

Eva L. Baker, University of California, Los Angeles, Center for Research on Evaluation, Standards, & Student Testing (CRESST), School of Education & Information Studies (SE&IS)

Randy E. Bennett, ETS, Research Institute

Jill Burstein, Duolingo, Inc.

Pamela Cantor, The Human Potential L.A.B.

Paul Cobb, Vanderbilt University

Linda Darling-Hammond, Learning Policy Institute

Jacqueline Darvin, Queens College, City University of New York

Ravit Dotan, TechBetter LLC

Kadriye Ercikan, Educational Testing Service

Carla M. Evans, National Center for the Improvement of Educational Assessment

Howard T. Everson, Graduate Center, City University of New York

Kate Felsen, The Human Potential L.A.B.

Natalie Foster, Organisation for Economic Cooperation and Development (OECD)

Sheryl L. Gómez, The Study Group

Edmund W. Gordon, Teachers College, Columbia University (Emeritus); Yale University (Emeritus)

Laura S. Hamilton, National Center for the Improvement of Educational Assessment.

John Hattie, University of Melbourne (Emeritus)

Kristen Huff, Curriculum Associates

Gerunda B. Hughes, Howard University (Emeritus)

Geoffrey T. LaFlair, Duolingo, Inc.

Carol D. Lee, Northwestern University (Emeritus)

Paul G. LeMahieu, Carnegie Foundation for the Advancement of Teaching; University of Hawai'i, Mānoa

Volume Contributors (continued)

Richard M. Lerner, Tufts University

Temple S. Lovelace, Assessment for Good, Advanced Education Research and Development Fund (AERDF)

Scott F. Marion, National Center for the Improvement of Educational Assessment

Maxine McKinney de Royston, Erikson Institute

Orrin T. Murray, Wallis Research Group

Na'ilah Suad Nasir, Spencer Foundation

Roy Pea, Stanford University

James W. Pellegrino, University of Illinois Chicago

Mario Piacentini, Organisation for Economic Cooperation and Development (OECD)

Stephen G. Sireci, University of Massachusetts Amherst, Center for Educational Assessment

Eric M. Tucker, The Study Group

Alina A. von Davier, Duolingo, Inc.

Kevin Yancey, Duolingo, Inc.

Credits

We gratefully acknowledge the leadership and dedication of the editorial team, whose vision, commitment, and expertise made the Handbook for Assessment in the Service of Learning series possible.

Series Editors
Edmund W. Gordon
Stephen G. Sireci
Eleanor Armour-Thomas
Eva L. Baker
Howard T. Everson
Eric M. Tucker

Managing Editors
Eric M. Tucker
Sheryl L. Gómez

We owe a profound debt of gratitude to *Professor Edmund W. Gordon* for his visionary conceptual leadership, which provided the inspiration and foundation for this Series. His friendship and decades-long commitment to scholarship that advances understanding of assessment in the service of learning has been the fountainhead throughout this project.

We gratefully acknowledge the *Gordon Seminar for Assessment in the Service of Learning,* housed at the Edmund W. Gordon Institute for Advanced Study at Teachers College, for its pivotal role in supporting the initial conceptualization of this Handbook. Convened by Professor Gordon starting in 2020 to advance the charge of the Gordon Commission for the Future of Assessment in Education, the Seminar provided a critical forum in which many of the ideas in these volumes were presented, debated, and refined. For over fifty years, the Gordon Institute has used advocacy, demonstration, evaluation, information dissemination, research and technical assistance to study and seek to improve the quality of life chances of communities of color through education in urban contexts.

We acknowledge *The Study Group* for stewarding the project to publication, including by assuming the project lead and managing editorial functions. The Study Group coordinated the solicitation and review of chapters, managed author communications, oversaw the copyediting, layout, and design, and delivered the manuscripts to the publisher. This leadership was essential to the Handbook's successful completion.

Acknowledgements

The *Handbook for Assessment in the Service of Learning* is the product of a dedicated community of scholars and practitioners, but we owe our most profound debt of gratitude to Professor Edmund W. Gordon. His scholarship provides the foundational inspiration and ethical compass for this series. From inception, Professor Gordon contributed the precious heirloom seed concepts planted and cultivated into the Handbook chapters. As convener of the Gordon Seminar for Assessment in the Service of Learning, housed at Teachers College, Columbia University, he fostered the rigorous inquiry and in-depth discussions that strengthened the core ideas forming the intellectual bedrock of these volumes. He challenged us to be ambitious, and his guidance was the essential element that sustained this collaboration. This Handbook series would not exist without him, and we are honored to carry forward his legacy.

We extend our sincere thanks to the Series Editors—Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, and Eric M. Tucker—whose collective vision, expertise, and commitment were instrumental in shaping the intellectual direction of this series. The Volume Co-Editors were fundamental in securing the quality of the scholarship within each volume. Guiding the operational and logistical dimensions of this complex process were our Managing Editors, Eric M. Tucker and Sheryl L. Gómez, who earned our thanks for their remarkable efforts in steering the processes from start to finish.

The conceptual origins of this Handbook series are rooted in the seminal work of the *Gordon Commission on the Future of Assessment in Education* (2011–2013). The Commission planted seeds that are beginning to come to fruition in these volumes. The Edmund W. Gordon Institute for Advanced Study in Education at Teachers College, Columbia University, now embracing its fifth decade, has served as the vital intellectual home for Professor Gordon and the ambitious projects he undertakes. We thank the Gordon Institute for the support that makes Professor Gordon's prolific scholarly life possible during his 104th year. We are grateful to Ezekiel Dixon-Román, the Gordon Institute's Director, and Paola Heincke, who have been steadfast partners for Professor Gordon's vision. We extend special thanks to Jonthon Coulson. His intellect, writing, curiosity, sense of adventure, and kindness

left an indelible mark on the Seminar and this Handbook, and we are particularly grateful for his stewardship and foundational organizational and conceptual contributions during the program's formative iterations.

The Gordon Seminar for Assessment in the Service of Learning formed a community of inquiry. The thoughtful feedback from its participants provided the intellectual space to test, develop, and strengthen the core ideas in these volumes. We offer profound appreciation to the Seminar's core participants: Eleanor Armour-Thomas, Aneesha Badrinarayan, Eva L. Baker, Randy Bennett, Susan M. Brookhart, Greg Chung, Madhabi Chatterji, Jonthon Coulson, Linda Darling-Hammond, Ezekiel J. Dixon-Román, Richard Durán, Howard T. Everson, Sheryl L. Gómez, Edmund W. Gordon, Kris D. Gutiérrez, Kenji Hakuta, Gerunda B. Hughes, Neal Kingston, Carol D. Lee, John Lee, Paul G. LeMahieu, Pamela Moss, Temple Lovelace, Susan Lyons, Robert J. Mislevy, Maria Elena Oliveri, Roy Pea, Jennifer Randall, Stephen G. Sireci, Eric M. Tucker, Ernest Washington.

We also thank the many distinguished colleagues who, as Seminar presenters and guests, challenged our assumptions and enriched our dialogue with their cutting-edge research: Itzel Aceves, Ryan Baker, Yoav Bergner, Abby Benedetto, John Behrens, Lauren Bierbaum, Jill Burstein, Tony Bryk, Pamela Cantor, Andy Calkins, Auditi Chakravarty, Andrew Dalton, Jacqueline Darvin, Kristen DiCerbo, Fabienne Doucet, Kadriye Ercikan, Dave Escoffery, Tianying (Teanna) Feng, Natalie Foster, Peter Gault, Jim Gee, E. Wyatt Gordon, Sunil Gunderia, Khaled J. Ismail, Fiona Hinds, Kristen Huff, JoAnn Hsueh, Neil T. Heffernan, Elizabeth Mokyr Horner, Rebecca Kockler, Timothy Knowles, Michael Kearns, Jade Caines, Richard Lerner, Lydia Liu, Maxine McKinney de Royston, Orrin Murray, Jasmine McBeath Nation, Britt Neuhaus, Osarugue Michelle Odemwingie, Andreas Oranje, Trevor Packer, Luciana Parisi, James W. Pellegrino, Bill Penuel, Mario Piacentini, Ramona Pierson, Elizabeth Redman, Jeremy Roberts, Barbara Rogoff, Maheen Sahoo, Amit Sevak, Lorrie A. Shepard, Laura Slover, Jim Shelton, Valerie Shute, Kim Smith, Rebecca Stone-Danahy, Natalya Tabony, Sylvane Vaccarino, Arthur VanderVeen, Alina von Davier, Alyssa Wise, Jason Yeatman. We are grateful to all who lent their expertise to this collaborative process, and we offer special thanks to Eleanor Armour-Thomas, Eric Tucker, and Sheryl Gómez for expertly moderating and organizing the Seminar. These sessions provided vital feedback on the Handbook chapters and framing as a work-in-progress.

We are thankful to the colleagues who participated in the AERA Honorary Presidential Sessions during annual meetings of the American Educational Research Association, providing a crucial platform for engaging with a range of viewpoints. The participants included: Brenda Allen, C. Malik Boykin, M. C. Brown, E. Wyatt Gordon, Jessica Heppen, Gabriela Lopez, Jamie Olson McKee, James L. Moore III, Na'ilah Suad Nasir, Anne Marie Núñez, Roberto J. Rodríguez, Timothy E. Sams, Mark Schneider, Matthew Soldner, LaVerne Evans Srinivasan, Claude Steele, Erica N. Walker, Amy Stuart Wells, Lester W. Young, Jr., and Elham Zandvakili.

Furthermore, a series of academic sessions convened to honor Professor Gordon's 100th birthday and his extraordinary legacy proved essential to this project's development. We extend our sincere gratitude to the host institutions, including Teachers College, Columbia University; University of California, Los Angeles; University of California, Santa Barbara; University of Massachusetts Amherst; and University of Texas at Austin. We thank the organizers and participants of these conferences; their engagement helped sharpen this Handbook series.

At the heart of this project are the contributions of the nearly 90 chapter authors whose collective scholarship forms the core of the Handbook. We thank them for their expertise and commitment. We are profoundly grateful to the series and volume editors, and peer reviewers, whose insightful critiques strengthened the quality, clarity, and coherence of each chapter.

We acknowledge The Study Group for its indispensable role in stewarding this project from conception to publication. Eric M. Tucker, Sheryl L. Gómez, Lauren Cutuli, and their team, expertly coordinated the complex processes of author communication, manuscript preparation and review, and production with skill and dedication. We are grateful to the University of Massachusetts Amherst Library for their partnership and commitment to open-access scholarship. The design and production of the Study Group and Dezudio design teams transformed our manuscripts into a polished and accessible final publication. This includes Ian Boly, Melissa Neely, Klaus Bellon, Ashley Deal, and Raelynn O'Leary.

We dedicate this Handbook to the memory of our cherished colleagues from the Gordon Commission who passed away during this work: Jamal Abedi, Lloyd Bond, A. Wade Boykin, Carl F. Kaestle, James Greeno, Stafford Hood, Robert J. Mislevy, and Lee Shulman. Their wisdom, friendship, and spirit were foundational

to this project, and their loss is deeply felt. We also remember all others from our community who have passed on; their contributions are woven into the fabric of this work, and we honor them with gratitude and respect.

Finally, on a personal note, we thank our families and friends for their support and patience throughout this journey. Our loved ones' understanding and encouragement sustained us through the long hours of research, writing, and editing. Each of the editors is grateful to those mentors and colleagues who offered personal support and guidance along the way—while too numerous to name here, please know that your influence has been invaluable.

In closing, we view the *Handbook for Assessment in the Service of Learning* as the harvest of many years of collaborative effort—a harvest that we are delighted to share with the world. Professor Gordon used an agricultural metaphor to describe this project, speaking of selecting and sowing conceptual seeds, cultivating fields, harvesting and milling wheat, and ultimately "breaking bread" together from the yield. Now, as these volumes go to press, it is nearly time to break bread in celebration of what has been achieved. We look forward to gathering—in person or in spirit—to enjoy and celebrate the harvest of ideas represented here. To everyone who has journeyed with us in bringing this Handbook series to fruition, thank you. We hope that the work born of this collective effort will, in turn, nourish further inquiry and innovation in the service of learning for generations to come.

Toward Assessment in the Service of Learning

Edmund W. Gordon

Pedagogical sciences and practice have long utilized educational assessment and measurement too narrowly. While we have leveraged the capacity of these technologies and approaches to monitor progress, take stock, measure readiness, and hold accountable, we have neglected their capacity to facilitate the cultivation of ability; to transform interests and engagement into developed ability. Assessment can be used to appraise affective, behavioral, and cognitive competence. From its use in educational games and immersive experiences, we are discovering that it can be used to enhance learning. Assessment, as a pedagogical approach, can be used to take stock of or to catalyze the development of Intellective Competence. Educational assessment as an essential component of pedagogy, in the service of learning, can inform and improve human learning and development. This Handbook, in three volumes, points us in that direction.

More than sixty years ago, I had the privilege of working alongside a remarkable educator, Else Haeussermann, whose insights into the learning potential of children with neurological impairments forever altered my understanding of educational assessment. At a time when many viewed such children as unreachable or incapable, Haeussermann insisted that their performances must be interpreted not merely to sort or classify, but to understand—and that understanding must inform instruction. Rather than measuring fixed abilities, she sought to uncover the conditions under which each child might succeed. Her lesson plans were not dictated by standardized norms, but by rich clinical observations of how learners engaged with tasks, responded to guidance, and revealed their ways of thinking. Though her methods defied the conventions of test standardization and were deemed too labor-intensive by prevailing authorities, they represented a

foundational model of what I now describe as *assessment in the service of learning;* assessment not as an endpoint, but as a pedagogical transaction—designed to inform, inspire, and improve the very processes of teaching and learning it seeks to illuminate. The lesson I took from Haeussermann was simple yet profound: that assessment should be used not only to identify what is, but to imagine and cultivate what might become. In every learner's struggle, there is the seed of possibility, and our charge as educators is to create the conditions under which that possibility can take root and flourish.

A Vision for Assessment in Education

In recent years, a profound shift has been gathering momentum in educational thought: the recognition that assessment should **serve** and **inform** teaching and learning processes–not merely measure their outcomes. Nowhere was this vision articulated more forcibly than by the Gordon Commission on the Future of Assessment in Education. Convened over a decade ago under my leadership, the Commission argued that traditional testing–focused on ranking students and certifying "what is"–must give way to new approaches that also illuminate how learning happens and how it can be improved. The Commission's technical report, *"To Assess, To Teach, To Learn"* (2013), proposed a future in which assessment is not an isolated audit of achievement, but rather a vital, integrated component of teaching and learning processes. It envisioned assessment practices that help *cultivate* students' developing abilities and inform educators' pedagogical choices, thereby contributing to the very intellective development we seek to measure. This call to re-purpose assessment—to make assessment a means for educating, not just evaluating—sets the stage for the present Handbook series. Since 2020, I have convened a group of leading scholars to advance the Commission's central proposition with urgency and optimism: that educational assessment, in design and intent, must be reconceived "in the service of teaching and learning."

The need for this reorientation has only grown more pressing. Conventional assessments, from high-stakes tests to admissions exams, have long been designed primarily to determine the achieved status of a learner's knowledge and skills at a given point in time. Such assessments can tell us how much a student knows or whether they meet a benchmark, which may be useful for the purpose of accountability and certification. Yet this traditional paradigm reveals little about how students learn, why they succeed or struggle, and what might help

them grow further. As I have often observed, an assessment system geared only toward outcomes provides a point-in-time picture–a static snapshot of developed ability–but does not illuminate the dynamic processes by which learners become knowledgeable, skilled, and intellectively competent human beings. In effect, we have been evaluating the outputs of education while neglecting the processes of learning that produce those outcomes. The result is an underutilization of assessment's potential: its potential to guide teaching, to inspire students, and to support the cultivation of intellective competence–that is, the capacity and disposition to use knowledge and thinking skills to solve problems and adapt to new challenges. To fulfill the promise of education in a democratic society, we must reimagine assessment as a positive force within teaching-learning processes, one that supports intellectual development, identity formation, equity, and human flourishing, rather than as an external judgment passed upon learning after the fact.

From Measurement to Improvement: Re-Purposing Assessment

Moving toward assessment in the service of learning requires candid reflection on the limitations of our prevailing assessment practices. Decades of research in educational measurement have given us reliable methods to rank, sort, and certify student performance. These methods excel at answering questions like: *What has the student achieved? or How does this performance compare to a norm or standard?* Such information is not without value–it can inform policy decisions, signal where resources are needed, and hold systems accountable for outcomes. However, as we refocus on learners themselves, a different set of questions comes to the fore: How can we improve learning itself? How can assessment and instruction work together to help students learn more deeply and effectively? Traditional tests rarely speak to these questions. A test score might tell us that a learner struggled with a set of math problems, but not *why*–Was it a misunderstanding of concept, a careless error, test anxiety, or something about the context of the problems? Nor does the score tell us *what next steps* would help the learner progress. In short, status-focused assessments alone do little to guide improvement. They measure the ends of learning but not the means.

By contrast, the vision of assessment espoused by the Gordon Commission and echoed in my volume "*The Testing and Learning Revolution*" (2015) is profoundly educative in its purpose. In this view, assessment is not a mere endpoint; it is part of an ongoing process of feedback and growth. When assessment is woven

into learning, it can provide timely insights to teachers and learners, diagnose misunderstandings, and suggest fruitful paths for further inquiry. It becomes a continuous conversation about learning, rather than a one-time verdict. This shift entails treating assessment, teaching, and learning as inseparable and interactive components of education—a dynamic system of influence and feedback. I describe assessment, teaching, and learning as a kind of troika or three-legged stool: each element supports and strengthens the others, and none should function independently of the whole. A test or quiz is not an isolated exercise; it is a transaction between the student, the educator, and the content, one that can spark reflection, adjustment, and new understanding. In this transactional view, the student is not a passive object of measurement but an active agent in the assessment process. How a learner interprets a question, attempts a task, uses feedback, or perseveres through difficulty—all of these are integral to the learning experience. Assessment tasks thus have a dual character: they both measure learning and simultaneously influence it.

Embracing this dual character opens up exciting possibilities for re-purposing assessment. Consider, for example, the power of a well-crafted problem-solving task. When a student grapples with a complex problem, the experience can trigger new reasoning strategies, reveal gaps in understanding, and ultimately lead to cognitive growth—if the student receives appropriate guidance and feedback. The late cognitive psychologist Reuven Feuerstein demonstrated decades ago that targeted "instrumental enrichment" tasks could significantly improve learners' thinking abilities; importantly, these tasks functioned as assessments and interventions at once. In the same spirit, assessments can be designed as learning opportunities: rich problems, projects, or simulations that *both* challenge students to apply their knowledge *and* teach them something in the process. A challenging science investigation, for instance, might double as an assessment of inquiry skills and a chance for students to refine their experimental reasoning. When students receive scaffolded support (hints, feedback, opportunities to try again), the assessment itself contributes to their development. In this way, assessment becomes a catalyst for learning. It shifts from a static checkpoint to a dynamic, educative experience. Each assessment interaction is an occasion for growth, not just an audit of prior learning.

Re-purposing assessment also calls for expanding the evidence we consider and collect about learning. If our aim is to understand learners' thinking and guide their progress, we must look beyond right-or-wrong answers. We need to examine *process:* How did the student arrive at this answer? What misconceptions were revealed in their intermediate steps? How did they respond to hints or setbacks? Such evidence may be gleaned through clinical interviews, think-aloud protocols, interactive tasks, or educational games that log students' actions. Today's technology makes it increasingly feasible to capture these rich process data. For example, a computer-based math puzzle can record each attempt a student makes, how long they spend, which errors they make, and whether they improve after feedback–yielding a detailed picture of learning in action. An assessment truly "in the service of learning" will tap into this kind of information, using it to formulate next steps for instruction and to provide learners with nuanced feedback on their strategies and progress. In short, we must broaden our view of what counts as valuable assessment data, integrating qualitative insights with quantitative scores to understand and support each learner's journey fully.

Assessment, Teaching, and Learning as Dynamic Transactions

Central to my proposed paradigm is the understanding that assessment is fundamentally relational and contextual. Learning does not unfold in a vacuum, and neither should assessment. Every assessment occurs in a context–a classroom, a culture, a relationship–and these contexts influence how students perform and how they interpret the meaning of the assessment itself. I speak of the "dialectical" relationship among assessment, teaching, and learning. By this they mean that these processes continuously interact and shape one another like an ongoing dialogue. A teacher's instructional move can be seen as a kind of assessment (gauging student reaction), just as a student's attempt on an assessment task is an act of learning and an opportunity for teaching. When we recognize this, assessment ceases to be a one-way transmission (tester questions, student answers) and becomes a two-way exchange–a *transaction*. In this transaction, students are active participants, bringing their own thoughts, feelings, and identities into the interaction. They are not simply responding to neutral prompts; they are also interpreting *what the assessment asks of them* and *why it matters.* In essence, assessment is a conversation about learning, one that should engage students as whole persons.

This perspective urges us to design assessments that are embedded in meaningful activity and closely tied to curriculum and instruction. Instead of pulling students out of learning to test them, the assessment becomes an organic part of the learning activity. For instance, a classroom debate can serve as an assessment of argumentation skills while also providing students with cycles of preparation and feedback regarding how to formulate and defend ideas. A collaborative applied research project can function as an assessment of problem-solving and teamwork, at the same time building those very skills through practice. In such cases, assessment and instruction intermingle; feedback is immediate and natural (peers responding to an argument, a teacher coaching during the project), and students often find the experience more engaging and relevant. The transactional view also highlights the role of relationships and identity in assessment. How a learner perceives the *purpose* of an assessment and their *relationship* to the person or system administering it will affect their engagement. Do they see the test as a threat or as an opportunity? Do they trust that it is fair and meant to help them? These factors can influence performance as much as content knowledge. Therefore, assessment in the service of learning must be implemented in a supportive, trustful environment. It should feel *to the student* like an extension of teaching—another way the teacher (or system) is helping them learn—rather than a judgment from on high. This more humane and dialogic approach aligns with my lifelong emphasis on humanistic pedagogy: education that honors the whole learner, respects their background and identity, and seeks to empower rather than stigmatize.

Embracing Human Variance and Equity

A commitment to humanistic, learner-centered assessment inevitably leads us to confront the reality of human variance. Learners differ widely in their developmental pathways, cultural and linguistic backgrounds, interests, and approaches to learning. I have often described human variance not as a complication to be managed, but as a core consideration and asset in education. Traditional standardized assessments, in their quest for uniform measures, have often treated variance as "noise" to be controlled or minimized. In contrast, assessment in the service of learning treats variation as richness to be understood and leveraged. Every learner brings a unique profile of strengths and challenges; a truly educative assessment approach seeks to personalize feedback and support to those individual needs. This is not only a matter of effectiveness but of equity

and justice. When assessment is used purely as a high-stakes gatekeeper, it has often exacerbated social inequalities–for example, by privileging those who are test-savvy or whose cultural background aligns with the test assumptions, while penalizing others with equal potential who happen to learn or express their knowledge in different ways. By re-purposing assessments to guide learning, we can instead strive to lift up every learner. Each student, whether gifted or struggling, whether English is their first or third language, whether learning in a suburban school or a remote village, deserves assessments that *help them grow.*

To achieve this, assessments must become more adaptive and culturally sustaining. They should be able to accommodate different ways of demonstrating learning and provide entry points for learners of varying skill levels (the idea of "low-floor, high-ceiling" tasks). They should also be sensitive to the cultural contexts students bring: the languages they speak, the values and prior knowledge they hold, the identities they are forming. An assessment that allows a bilingual student to draw on both languages, for instance, may better capture–and cultivate–that student's full communicative ability. Similarly, assessments can be designed to honor diverse knowledge systems and ways of reasoning, rather than only a narrow canon. When students see their own experiences and communities reflected in what is being assessed, they are more likely to find meaning and motivation in the task. Moreover, such inclusive assessments can play a role in identity formation: they send a message to students about what is valued in education and whether *they* belong. If assessments primarily signal to some students that they are "failures" or "deficient," those students may internalize negative academic identities, which can undermine their confidence and engagement. But if assessments are reimagined to recognize growth, effort, and multiple and varied abilities, students can begin to see themselves as capable, evolving learners. In this way, a re-purposed assessment system supports not only cognitive development but also the formation of a positive learner identity for every student. Ultimately, embracing human variance is crucial to realizing the broader aim of human flourishing. Education is about nurturing the potential of each human being; assessment should be an instrument for that nurture, helping all learners discover and develop their capabilities to the fullest.

Toward a Pedagogical Renaissance: Analytics and Intellective Competence

Realizing the vision of assessment in the service of learning will require innovation and a renewed research agenda–what we might call a pedagogical renaissance in assessment. One promising path I have begun to explore is the development of "pedagogical analyses" as a robust practice in education. Pedagogical analysis refers to the systematic study of how teaching, learning, and assessment interact–using all available data to understand what works for whom and why. With modern technology, we have more data than ever before about learners' interactions (click streams, response times, error patterns, etc.), and powerful analytical tools, including machine learning, to detect patterns in this data. The goal of pedagogical analysis is not mere number-crunching for its own sake, but to generate actionable insights into the learning process. For example, an analysis might reveal that a particular sequence of hints in an online tutoring system is especially effective for learners who initially struggle, or that students with specific background knowledge benefit from a different task format. These insights allow educators and assessment designers to refine their approaches, tailoring them to a wide range of learners–in essence, personalizing assessment and instruction on a large-scale. Importantly, this data-driven approach must be guided by sound theory and a humanistic compass: we seek not to reduce learners to data points, but to augment our understanding of their intellective competence and how it grows.

The concept of intellective competence is central here. Intellective competence, a term I coined, denotes the ability and disposition to use one's knowledge, strategies, and values to solve problems and to continue learning. It is a holistic notion of what it means to be an educated, capable person–going beyond the memorization of facts or routine skills. Our assessment systems should ultimately aim to foster *and* capture these broad competencies: critical thinking, adaptability, creativity, and the capacity to learn how to learn. Doing so means designing assessments that pose authentic, complex challenges to students and then analyzing not only whether students got answers correct, but *how* they approached the challenge. Did they show ingenuity in finding a solution? Did they learn from initial failures and try alternative strategies? Such qualities are the hallmarks of intellective growth. By gathering evidence of these behaviors, we align assessment with the real goals of education in the 21st-century. Moreover, assessing for intellective competence has the positive side effect of encouraging teaching toward deeper learning, rather than teaching to a narrow test. When assessments value reasoning, exploration, and

resilience, teachers are more likely to cultivate those capacities in their students. In this way, re-purposed assessments can help bring about a richer educational experience for learners–one that genuinely prepares them for lifelong learning and flourishing in a complex world.

Of course, moving from our current assessment paradigm to this envisioned future is a substantial endeavor. It raises important questions for policy, practice, and research. Policymakers will need to broaden accountability systems to value growth and process, not just point-in-time proficiency. Educators will need professional support to use formative assessment strategies effectively and to interpret the richer data that new assessments provide. Researchers must continue to investigate the best ways to design and implement assessments that embed learning, as well as develop valid ways to infer student understanding from interactive tasks and big data patterns. These challenges, while significant, are surmountable. Indeed, around the world we already see glimpses of the possible: innovative formative assessment programs that transform classrooms into collaborative learning labs; game-based assessments that engage children and teach new skills; participatory assessment approaches that involve students in self- and peer-evaluation, building their metacognitive awareness. Such examples are heartening "existence proofs" that assessment can be reimagined to the benefit of everyone. The task now is to build on these successes, knitting them into a coherent approach that can be implemented broadly and equitably.

The Journey Ahead–and the Contributions of this Handbook Series

This Handbook for Assessment in the Service of Learning series stands as a timely and essential contribution to this educational renaissance. Across its volumes, a breadth of perspectives is presented, all converging on the central theme of transforming assessment to better support teaching and learning. The chapters compiled here bring together renowned scholars and practitioners from a wide range of fields, including cognitive science, psychometrics, artificial intelligence, learning sciences, curriculum and learning design, educational technology, sociology of education, and more. Such range is intentional and necessary. Rethinking assessment is a complex endeavor that benefits from multiple lenses: theoretical, empirical, technological, and practical. Some contributions explore foundational theoretical frameworks, helping us reconceptualize what assessment is and *ought to be* in light of contemporary knowledge about how people learn.

Others delve into the design of innovative assessments, offering design principles and prototypes for assessments that measure complex competencies or integrate seamlessly with instruction. We also encounter rich case studies and practical exemplars–from early childhood settings to digital learning environments–that demonstrate how assessment for learning can be implemented on the ground. These range from classrooms where teachers have successfully used formative assessment to empower students, to large-scale programs that blend assessment with curriculum, to cutting-edge uses of data analytics and AI solutions that personalize learning experiences. The wide-ranging nature of these examples underscores a crucial point: *assessment in the service of learning is applicable in a significant range of educational contexts*. Whether in formal preK-12 schooling, higher education, workplace training, informal learning, or through media and games, the principles remain relevant–aligning assessment with growth, understanding, and human development.

While the chapters in this series each offer unique insights, they are united by a spirit of inquiry, urgency, and hope that echoes the ethos of the Gordon Commission. There is inquiry–a deep questioning of assumptions that have long been taken for granted, such as the separation of testing from teaching, or the notion that ability is a fixed trait to be measured. There is urgency–a recognition that as we move further into the 21st century, with its rapid social and technological changes, the costs of clinging to outdated assessment regimes are too great. We risk stifling creativity, perpetuating inequity, and mis-preparing learners for a world that demands adaptability and continuous learning. But above all, there is hope–a belief that through thoughtful innovation and collaboration, we *can* redesign assessment to be a positive force in education. The work is already underway, and this Handbook is part of it. The range of perspectives in these volumes is a source of strength, encompassing critical analyses, bold experiments, and a blend of longstanding wisdom and fresh ideas, each contributing a piece to the larger puzzle of how to make assessment truly *for* learning.

In closing, let us return to the animating vision that I have championed throughout my career and which inspires this series. It is a vision of education where every learner is seen, supported, and challenged; where assessment is not a grim rite of ranking, but a continuous source of insight and improvement; where teaching, learning, and assessment form a holistic enterprise devoted to nurturing the growth of human potential. Realizing this vision will require perseverance and

creativity. It will mean overcoming institutional inertia and reimagining roles–for test-makers, teachers, students, and policymakers alike. Yet the potential payoff is immense. By making assessment a partner in learning, we stand to enrich the educational experience for all students, help teachers teach more effectively, and advance the cause of equity and excellence by ensuring that every learner receives the feedback and opportunities they need to thrive. This is assessment in the service of learning: assessment that not only reflects where learners are, but actively helps them get to where they need to go next. With the insights and evidence gathered in this Handbook series, we take important steps on that journey. The message is clear and hopeful–it is time to move beyond the extant paradigm and embrace a future in which to assess is, intrinsically, to teach and to learn.

References

The Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment* (Technical report). Educational Testing Service. https://www.ets.org/Media/Research/pdf/gordon_commission_technical_report.pdf

Gordon, E. W., & Rajagopalan, K. (2016). *The testing and learning revolution: The future of assessment in education.* Palgrave Macmillan US. https://doi.org/10.1057/9781137519962

Handbook for Assessment in the Service of Learning Series Preface

Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, and Eric M. Tucker

Objective

How might educational assessment become a catalyst for learning and human development? This question lies at the heart of the *Handbook for Assessment in the Service of Learning* series, Volumes I, II & III. This series provides a research-based introduction to the theory, design, and practice of assessment in the service of teaching and learning (Gordon, 2020; 2025). The Handbook echoes the call of the *Gordon Commission on the Future of Assessment in Education* to repurpose assessment from merely certifying 'what is' to illuminating how learning happens and how it can be improved (Gordon Commission, 2013; Gordon, 2025). The three volumes presented here respond to that call.

Description

The three volumes in this series offer a contemporary view of a range of theoretical perspectives, scholarship, and research and development on innovations with the potential to enable assessment to enhance learning. Across the volumes, contributors explore the central theme of transforming assessment design and development to better support teaching and learning. The three volumes draw on the sciences of learning, measurement, pedagogy, improvement, and more–to inform this charge. We asked authors to anchor chapters in one or more of the design principles for assessment in the service of learning (Baker, Everson, Tucker, & Gordon, 2025). The chapters probe longstanding assumptions, and they explore how to weave a focus on learning into the fabric of educational assessments. The interested reader will find working examples that illustrate what these emerging approaches might look like in practical contexts, from classroom assessments that empower student agency, to larger-scale assessment systems that, by design,

integrate with curriculum and instruction, to applications of data analytics and AI-powered learning platforms that personalize assessment and promote learning. Together, these contributions reflect a common inquiry regarding the design, development, and use of assessment not merely to certify what students know and can do, but to illuminate and support how learning happens and can improve, for every learner (Gordon, 2025; Gordon & Rajagopalan, 2016; Shepard, 2019). From the learner's perspective, well-crafted assessments catalyze and cultivate the very understanding and performance they elicit. Accordingly, the goal is to design educational assessments to nurture productive struggle and growth in the learner.

Audience

This Handbook is intended for a broad audience, from test developers, assessment researchers, and learning scientists to educators, policy makers, and designers. It is a resource for anyone interested in using assessment to help learners learn.

Organization

This Handbook for Assessment in the Service of Learning series is organized into three volumes, each focusing on a critical dimension of assessment in the service of learning. The series includes:

- Volume I: Foundations for Assessment in the Service of Learning
- Volume II: Reconceptualizing Assessment to Improve Learning
- Volume III: Examples of Assessment in the Service of Learning

Together, the volumes present a holistic picture of what it means to redesign assessment in the service of learning–from high-level design frameworks down to concrete tools and practices, and from classroom-level interventions to system-wide exemplars.

Rationale

Too often, assessments have been treated as end-of-learning verdicts–snapshots of what students have achieved–rather than as integral parts of the learning process (Pellegrino, 2014). Meanwhile, important domains of student ability (complex skills like critical thinking and collaboration) have been poorly captured by conventional tests that focus narrowly on easily measured skills (Gordon, 2020).

This Handbook responds to Gordon's charge for assessment innovation. By showcasing successful exemplars, these volumes help define and shape the field that has emerged in the years since the Gordon Commission. Assessment in the service of learning represents a shift in perspective that views assessment, teaching, and learning as inseparable, entangled processes. It envisions a future where every learner is understood, appropriately supported, and sufficiently challenged (Gordon, 1996; Goldman & Lee, 2024). When assessment becomes a partner in the pedagogical aspects of curriculum and instruction, it can enrich and improve teaching and help every learner thrive (Armour-Thomas & Gordon, 2025; Hattie, 2009; Ruiz-Primo & Furtak, 2024). This is the promise of assessment in the service of learning: to not only reflect where learners are, but to actively help them get to where they need to go next. The message of this Handbook is clear: it is time to embrace a future where to assess is to teach and to learn.

References

Armour-Thomas, E., & Gordon, E. W. (2025). *Principles of dynamic pedagogy: An integrative model of curriculum, instruction, and assessment for prospective and in-service teachers.* Routledge.

Baker, E. L., Everson, H., Tucker, E. M., & Gordon, E. W. (2025). Principles for assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Goldman, S. R., & Lee, C. D. (2024). Human learning and development: Theoretical perspectives to inform assessment systems. In S. F. Marion, J. W. Pellegrino, & A. I. Berman (Eds.), *Reimagining balanced assessment systems* (pp. 48–92). National Academy of Education.

Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment: Technical Report.* Educational Testing Service.

Gordon, E. W. (2020). Toward assessment in the service of learning. Educational Measurement: *Issues and Practice, 39*(3), 72–78.

Gordon, E. W. (2025). Series introduction: Toward assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Gordon, E. W., & Rajagopalan, K. (2016). *The Testing and Learning Revolution: The Future of Assessment in Education* (pp. 107–146). New York: Palgrave Macmillan US.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Routledge.

Pellegrino, J. (2014). Assessment in the service of teaching and learning: Changes in practice enabled by recommended changes in policy. *Teachers College Record, 116* (110313). https://doi.org/10.1177/016146811411601102

Ruiz-Primo, M. A., & Furtak, E. M. (2024). Classroom activity systems to support ambitious teaching and assessment. In S. F. Marion, J. W. Pellegrino, & A. I. Berman (Eds.), *Reimagining balanced assessment systems* (pp. 93–131). National Academy of Education.

Shepard, L. A. (2019). Classroom assessment to support teaching and learning. *The Annals of the American Academy of Political and Social Science, 683*(1), 183–200. https://doi.org/10.1177/0002716219843818

Foundational Issues of Assessment in the Service of Learning

Eric M. Tucker and Eleanor Armour-Thomas

In his introduction to this Handbook for Assessment in the Service of Learning series, Professor Edmund W. Gordon challenges us to reimagine educational assessment as a pedagogical act rather than a mere measurement exercise (Gordon, 2025). Traditional testing—long focused on ranking students and certifying "what is"—must give way to new approaches that illuminate how learning and development happen and how those processes can be improved (Armour-Thomas & Gordon, 2025; Armour-Thomas, 2025; Gordon, 2020, 2025; Gordon Commission on the Future of Assessment in Education, 2013). Volume I outlines design principles and technologies for assessment systems that both measure and foster learning, attentive to learner identities and contexts (Gordon, 2025).

Why Foundations, Why Now?

Volume I, *Foundations for Assessment in the Service of Learning*, makes the case for rethinking assessment design. First, assessment should generate insights on what matters most rather than merely what is easy to measure. This means broadening the constructs we assess to include the complex cognitive, affective, and behavioral skills needed for success in the 21st-century (Huff, 2025; National Research Council, 2001). Second, to assess these outcomes and processes, we need new design approaches that leverage learning sciences and frontier technologies. The affordances of emerging technology—from interactive simulations to generative artificial intelligence and data analytics—allow us to create assessments that provide more authentic tasks and timely feedback, blurring assessment and instruction (Foster & Piacentini, 2025). Third, innovation in assessment in the service of learning must center on the "value proposition"; the usefulness and usability of insights and interpretation generated for learners, educators, and families (Pellegrino, 2025). We must ensure that new assessments yield valid, trustworthy evidence of student learning and do so fairly across a wide range of learners. We must expand *what* we assess (in terms of affective,

behavioral, and cognitive skills), *how* we generate evidence, while also upholding quality with regard to *how well* assessment provides value through high-quality insights for learners and educators (Haertel, 2013; Messick, 1989; Moss, Girard, & Haniford, 2006). Educators and policymakers recognize that assessment *for* and *as* learning must complement assessment *of* learning; the paradigm of one-size-fits-all testing is simply inadequate for guiding improvement or promoting deeper understanding.

Contributors to Volume I recognize that assessment has historically been a double-edged sword: while it can empower learners and improve teaching, it also has been used to sort, gatekeep, or harm marginalized students (Gordon, 1996; Penuel & Watkins, 2019). Centering the needs and assets of students who have been least well served by traditional tests. A learner-centered approach also calls for involving the primary *users* of assessment in their development: teachers, learners, and families should help decide what is assessed and how results are used, so that assessment becomes collaborative and democratic (Chatterji, 2025). Baker and colleagues identify core principles to inform the building of assessments in the service of learning—including clarity of purpose, transparency, equity, amplifying learning, and quality assurance—so that assessment, in any subject, is deliberately crafted to support learning for every student (Baker, Everson, Tucker, & Gordon, 2025). In sum, Volume I is built on three foundational issues about assessment in the service of learning; (1) Design principles and frameworks for assessment in the service of learning; (2) Theoretical and empirical ideas regarding how people learn and what it means for assessment; and (3) Technological innovations reshaping educational assessment.

Section I–Design Principles and Frameworks for Assessment as Learning

Section I introduces foundational design principles and frameworks for reimagining assessment as an integral part of teaching and learning, and addresses how assessment systems can be designed to effectively support learner variation and improve learning. Additionally, the chapters call for transformative changes at the system level, advocating responsive, relevant, human-centered designs (Badrinarayan, 2025; Lovelace, Murray, & Hamilton, 2025). It emphasizes the need to leverage emerging technologies, maintain principled design methodologies, and align approaches for scalable innovation (Hattie, Sireci, & Baker, 2025).

Section II–Research Foundations: How People Learn and What It Means for Assessment

Section II establishes a research-driven foundation for transforming assessment, anchored in contemporary understandings of how people learn and develop across contexts (Lerner & Cantor, 2025; Pea, Lee, Nasir, & McKinney DeRoyston, 2025). It underscores the need for assessments aligned with the complex, situated nature of learning. A primary theme is aligning assessment practices with the complexity of human learning, recognizing that learners develop through interactions with families, communities, and schools. Effective assessment must capture this dynamic interplay, reflecting not only learners' immediate knowledge but also the conditions and supports influencing their ongoing development (Armour-Thomas, Darvin, & Hughes, 2025). Section II underscores the need for culturally responsive assessments, proposing tasks and methods to honor learners' cultural assets and ways of knowing while maintaining rigor (Badrinarayan, Bennett, & Darling-Hammond, 2025).

Section III–Emerging Innovations and New Possibilities

Section III explores emerging innovations and technologies, particularly artificial intelligence (AI) and advancing digital and data tool affordances, to reshape educational assessment. This section emphasizes how these technologies might enable a deeper, more nuanced understanding of student learning through simulations, adaptive tasks, and rich learner data. This section examines essential challenges posed by these advancements, including ensuring validity, reliability, fairness, and ethical integrity of AI-enabled assessment practices (Ercikan, 2025; Burstein, LaFlair, Yancey, von Davier, & Dotan, 2025). The chapters underscore that meaningful integration of innovation requires thoughtful alignment of technological capabilities with a steadfast ethical commitment, ensuring assessment remains human-centered, equitable, and supportive of deeper learning.

References

Armour-Thomas, E. (2025). Dynamic pedagogy: A perspective for integrating curriculum, instruction, and assessment in the service of learning at the classroom level. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries Libraries.

Armour-Thomas, E., & Gordon, E. W. (2025). *Principles of dynamic pedagogy: An integrative model of curriculum, instruction, and assessment for prospective and in-service teachers.* Routledge.

Armour-Thomas, E., Darvin, J., & Hughes, G. B. (2025). Assessment as a pillar of pedagogy in support of learning in AP Research and mathematics education courses. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries Libraries.

Badrinarayan, A. (2025). Reimagining state assessments in service of teaching and learning: Design principles for instructionally relevant assessments. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Badrinarayan, A., Bennett, R. E., & Darling-Hammond, L. (2025). Perspectives on socioculturally responsive assessment in large-scale systems. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Baker, E. L., Everson, H., Tucker, E. M., & Gordon, E. W. (2025). Principles for assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Burstein, J., LaFlair, G. T., Yancey, K., von Davier, A. A., & Dotan, R. (2025). Responsible AI for test equity and quality: The Duolingo English Test as a case study. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Chatterji, M. (2025). *User-centered assessment design: An integrated methodology for diverse populations.* The Guilford Press.

Ercikan, K. (2025). Efficacy, validity and fairness considerations in AI-driven assessments. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Foster, N., & Piacentini, M. (2025). Innovating assessment design to better measure and support learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment: Technical Report.* Educational Testing Service.

Gordon, E. W. (1996). Toward an equitable system of educational assessment. *The Journal of Negro Education, 64*(3), 360–372.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice, 39*(3), 72–78.

Gordon, E. W. (2025). Series introduction: Toward assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Haertel, E. (2013). Expanding views of interpretation/use arguments. *Measurement: Interdisciplinary Research and Perspectives, 11*(1–2), 68–70.

Hattie, J., Sireci, S. G., & Baker, E. L. (2025). Mind frames for improving educational assessment. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Huff, K. (2025). Designing and developing educational assessments for contemporary needs. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Lerner, R. M., & Cantor, P. (2025). Implications of a dynamic, relational-developmental-systems perspective for research design, measurement, and data analysis in the service of understanding and enhancing youth development and learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Lovelace, T. S., Murray, O. T., & Hamilton, L. S. (2025). Designing for the future: Toward an R&D agenda to promote inclusive, human-centered assessment systems. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education, 30*(1), 109–162.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academies Press.

Pea, R., Lee, C., Nasir, N., & McKinney DeRoyston, M. (2025). The cultural foundations of learning: Design considerations for measurement and assessment. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Pellegrino, J. W. (2025). Arguments in support of innovating assessments. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Penuel, W. R., & Watkins, D. A. (2019). Assessment to promote equity and epistemic justice: A use-case of a research-practice partnership in science education. *The Annals of the American Academy of Political and Social Science, 683*(1), 201–221.

VOLUME I | SECTION 1

Design Principles, Considerations, and Affordances for Assessment in the Service of Teaching and Learning

Measure What Matters for Teaching and Learning, and Measure It Well: Considering Design Principles and Approaches

Eric M. Tucker and Eleanor Armour-Thomas

Assessment must be reconceived as a tool for learning, not simply a tool for measurement (Baker & Gordon, 2014; Bennett, 2017; Shepard, 2019). For too long, tests have been used primarily to rank and sort students, reflecting aspects of what is—their current achievement—without guiding efforts to inform and improve what could be with regard to learning and development (Gordon, 1995, 2020, 2025; Gordon & Rajagopalan, 2016; Shepard, 2021). Volume I contributors argue that this paradigm must give way to approaches that illuminate how learning happens and how instruction can improve (Tucker & Armour-Thomas, 2025). If we hope to transform assessment systems, we must transform how we design, develop, and orient assessments–with clear purpose, strong grounding in how students learn, and a commitment to efficacy, quality, and usefulness in classrooms and learning contexts (Baker, Everson, Tucker, & Gordon, 2025; Huff, 2025; National Research Council, 2001; Pellegrino, 2025). The chapters in this section lay a foundation for human-centered and balanced assessment systems that foster deeper learning for every learner and address the real constraints that come with such a challenge.

Baker and colleagues (Baker et al., 2025) articulate design principles: be transparent about what is assessed, how, and to what ends; make purpose and focus explicit—clarifying intended outcomes, indicators of progress, and the learner processes that should transfer across contexts; and engineer tasks that support motivation, attention, and metacognition. They further urge designers to model trajectories over time; link assessment to feedback and adaptive instruction; accommodate learner variation with fair tasks and supports; and underwrite all claims with accessible evidence of quality and validity aligned to intended uses.

Pellegrino makes the case for assessment innovation, framing assessment as a process of reasoning from evidence and presents three arguments. First, we must "measure what matters, not just what is easy to measure" (Pellegrino, 2025). This means complex competencies—cognitive, socio-cognitive, and socio-emotional skills—that are essential for success in the future (Darling-Hammond, Herman, Pellegrino, Abedi, Aber, Baker, Bennett, Gordon, Haertel, Hakuta, Ho, Linn, Pearson, Popham, Resnick, Schoenfeld, Shavelson, Shepard, Shulman, & Steele, 2013; Foster & Piacentini, 2025). Second, we need design and delivery approaches that leverage modern technologies and tasks to capture richer evidence of learning (Huff, 2025; Linn, Baker, & Dunbar, 1991). Third, value proposition innovation in service of educators and learners is only as powerful as the evidence that underwrites it. We must apply rigorous standards of validation, reliability, and fairness to fulfill an obligation to add value to learners and educators (Chatterji, 2025; Pellegrino, 2025). In short, the promise of innovative assessment is inseparable from a robust evidentiary foundation; without trustworthy data, even the most imaginative assessment tasks are unlikely to serve learning.

Foster and Piacentini (2025) recognize that 21st-century education demands assessments of more complex skills—and that innovation is needed *across* the entire development process, from defining what to measure, to task design, scoring, and reporting (Darling-Hammond & Adamson, 2014; Marion, Pellegrino, & Berman, 2024). Piacentini and Foster propose five design principles for next-generation assessments, each grounded in research on how people learn. A student working through an extended task with built-in feedback is simultaneously being assessed *and* taught.

Lovelace, Murray, and Hamilton (2025) argue that transforming assessment requires rethinking the research and development (R&D) infrastructures. To counter system incoherence, anchor assessment in inclusive, human-centered design (Lovelace, Murray, & Hamilton, 2025). This means that assessment innovation should start with the needs and variation of learners at the center, honoring students' varied cultural and developmental contexts, and involve educators and communities in the design process. Lovelace and colleagues (2025) propose a three-layer agenda—modern, culturally relevant content; tools (e.g., simulations, AI-driven feedback); and enabling policy.

Huff (2025) observes that educational assessment stands at a "critical inflection point". The conventional approaches to test design that prevailed for decades, methods rooted in an era of norm-referenced tests used mainly to rank-order students, are no longer sufficient (Marion et al., 2024; Shepard, 2000). Today, there is growing demand for assessments that are "authentic, informative, actionable, engaging and accessible for all students" (Huff, 2025). First, the constructs we aim to measure have grown more complex–for example, mathematical practices, scientific inquiry skills, or collaborative problem-solving rich domains. Second, stakeholders expect assessments to serve multiple purposes (from guiding classroom teaching to informing accountability) without causing test overload, which means a single assessment design often must support several interpretations and uses. Third, there is heightened scrutiny of quality and fairness, pressuring developers to provide validity evidence and to design with rigor and transparency. Huff (2025) recommends Principled Assessment Design approaches, including evidence-centered design, to ensure task–claim coherence and explicit validity from the outset.

If we do not begin with design principles that privilege purpose, efficacy, coherence, balance, and evidence-based design, we risk innovating in name only–producing new solutions that replicate old problems. But as the chapters argue, if we design with integrity and intentionality, assessments have the potential to become powerful levers to improve and evaluate learning at all levels of the educational system—classroom, school, district, and state. The knowledge base on learning, cognition, and measurement has never been stronger, and the urgency for more effective, more humane assessment has never been clearer (Gordon, 2025; Mislevy, 2018; National Research Council, 2001; Penuel & Watkins, 2019). By centering design principles that prioritize learning, we–researchers, policymakers, educators, and assessment developers alike–can work together to build assessment systems that empower every student to learn and develop.

References

Baker, E. L., & Gordon, E. W. (2014). *From the assessment of education to the assessment for education: Policy and futures. Teachers College Record, 116,* 1–24.

Baker, E. L., Everson, H., Tucker, E. M., & Gordon, E. W. (2025). Principles for assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Bennett, J. (2017, December 8). *Assessment FOR learning vs. assessment OF learning.* Pearson. https://www.pearsonassessments.com/professional-assessments/blog-webinars/blog/2017/12/assessment-for-learning-vs--assessment-of-learning.html

Chatterji, M. (2025). *User-centered assessment design: An integrated methodology for diverse populations.* The Guilford Press.

Darling-Hammond, L., & Adamson, F. (2014). *Beyond the bubble test: How performance assessments support 21st-century learning.* Jossey-Bass.

Darling-Hammond, L., Herman, J., Pellegrino, J. W., Abedi, J., Aber, J. L., Baker, E., Bennett, R., Gordon, E., Haertel, E., Hakuta, K., Ho, A., Linn, R. L., Pearson, P. D., Popham, W. J., Resnick, L., Schoenfeld, A. H., Shavelson, R., Shepard, L. A., Shulman, L., & Steele, C. M. (2013). *Criteria for high-quality assessment.* Stanford Center for Opportunity Policy in Education.

Foster, N., & Piacentini, M. (2025). Innovating assessment design to better measure and support learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Gordon, E. W. (1995). Toward an equitable system of educational assessment. *The Journal of Negro Education, 64*(3), 360–372.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice, 39*(3), 72–78.

Gordon, E. W. (2025). Series introduction: Toward assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Gordon, E. W., & Rajagopalan, K. (2016). *The Testing and Learning Revolution: The Future of Assessment in Education* (pp. 107–146). New York: Palgrave Macmillan US.

Huff, K. (2025). Designing and developing educational assessments for contemporary needs. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

Lovelace, T. S., Murray, O. T., & Hamilton, L. S. (2025). Designing for the future: Toward an R&D agenda to promote inclusive, human-centered assessment systems. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Marion, S. F., Pellegrino, J. W., & Berman, A. I. (Eds.). (2024). *Reimagining balanced assessment systems.* National Academy of Education.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement.* Routledge.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* National Academies Press. https://doi.org/10.17226/10019

Penuel, W. R., & Watkins, D. A. (2019). Assessment to promote equity and epistemic justice: A use-case of a research-practice partnership in science education. *The Annals of the American Academy of Political and Social Science, 683*(1), 201–221. https://doi.org/10.1177/0002716219843249

Pellegrino, J. W. (2025). Arguments in support of innovating assessments. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Shepard, L. A. (2019). Classroom assessment to support teaching and learning. *The Annals of the American Academy of Political and Social Science, 683*(1), 183–200. https://doi.org/10.1177/0002716219843818

Shepard, L. A. (2021). Ambitious teaching and equitable assessment: A vision for prioritizing learning, not testing. *American Educator, 45*(3), 28–37, 48.

Tucker, E. M., & Armour-Thomas, E. (2025). Foundational Issues of Assessment in the Service of Learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Principles for Assessment in the Service of Learning

Eva L. Baker, Howard T. Everson, Eric M. Tucker, and Edmund W. Gordon

Abstract

This chapter offers a set of seven principles to guide the design and use of learning-focused assessments, that is, educational tests and assessments intended to support student learning. In the Handbook volumes, the principles were intended to assist chapter authors in considering these common elements in their contributions. In fact, some authors have chosen to include their own rendition of recommended principles. However, this brief chapter enumerates the principles developed by the participants of the Gordon Seminar. The principles are followed by a general rationale to support their inclusion as core components of an integrated design framework for assessments and concomitant learning and instruction. The seven principles are: (1) Assessment transparency; (2) Assessment focus and purpose; (3) Assessment support of learning processes (including attention, motivation, engagement, effort, and metacognition); (4) Assessment modeling of expectations and desired learning over time; (5) Assessment-linked instructional support including feedback; (6) Assessment equity with attention to learner variation; and (7) Assessment quality and validity involving the development of evidence to assure the value of assessment and learning to the learner and teacher.

Introduction

To guide the design, development, and use of assessments that are intended to support learning, we provide a set of principles to serve as the foundation for a conceptual framework for teachers, developers, and users of assessment results. Educational assessments with the deliberate intent of optimizing learning and teaching, as opposed to reporting student status, will have different characteristics from traditional large-scale assessments. Intended for use in any educational or training system, the proffered principles apply across academic disciplines and subject matter domains, settings, and ages and backgrounds of learners. The following principles form the crux of our recommendations.

- **Principle 1:** Assessment **transparency** provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.
- **Principle 2:** Assessment **focus** is explicit and includes purposes, outcomes, progress indicators, and processes that can be transferred to other settings, situations, and conditions.
- **Principle 3:** Assessment design supports learners' **processes**, such as motivation, attention, engagement, effort, and metacognition.
- **Principle 4:** Assessments model the structure of **expectations** and **desired learning** over time.
- **Principle 5: Feedback**, adaptation, and other relevant instruction should be linked to assessment experiences.
- **Principle 6:** Assessment **equity** requires fairness in design of tasks and their adaptation to permit their use with respondents of different backgrounds, knowledge, and experiences.
- **Principle 7: Assessment quality and validity** should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility.

Structure of Chapter

We provide the logic of the development of these principles in an overall rationale that includes the background of the Gordon Seminar for Assessment in the Service of Learning and the focus on assessment in support of learning. Next, we give an additional explanation relevant to the set of principles and briefly consider key topics in turn.

Finally, we offer a summary. Note that we have chosen not to include a significant number of footnotes or references within the text to allow the reader to make easy progress through this material, almost all of which will be expanded or illustrated in subsequent chapters. For those with interest in the scientific or experiential bases of the principles, we refer you to the selected bibliography at the conclusion of this narrative. For each principle, the selected bibliography provides a set of references that highlight its theoretical and empirical underpinnings.

A Rationale for Assessments Designed Explicitly to Inform Learning and Teaching

In the hands of skilled educators (e.g., teachers and developers), we believe that these principles will support a conceptual framework useful for designing educational assessments and tests to meet the needs of learners and of those engaged in helping people learn. The principles provide brief but concrete suggestions for repurposing and improving the design and use of educational assessments to enable students to improve their learning and performance in emerging educational contexts.

The background of these principles stemmed from the conversations of the Gordon Seminar: a weekly online meeting of scholars engaged in discussions for five years about how to follow up and operationalize the findings of *The Gordon Commission on the Future of Assessment in Education* (Gordon Commission on the Future of Assessment in Education, 2013) That publication was "created to consider the nature and content of American education during the 21st-century and how assessment can be used most effectively to advance that vision by serving the educational and informational needs of students, teachers, and society" (Pellegrino, 2014, p. 1).

The key message of the Gordon Commission underscored the importance of employing assessment to support the growth of learners in contrast to its dominant use to report student status in general, comparative formats. These status results almost always depend on annual or other periodic large-scale administrations of standardized tests. The test content of these examinations may be only loosely connected to the realities of classroom activities. For example, test developers with their administrative time constraints and psychometric requirements invariably choose to interpret curriculum standards or goals differently (and necessarily more generally) than individual classroom teachers.

In contrast to these uses of tests for status reporting practices, the Commission's Report emphasized that assessment and its findings should directly benefit student learning as well as mediate the efforts of teachers or other instructional designers to support their students. To be sure, scholars and practitioners have been advocating the use of assessment to support learning for decades, but the Commission pulled together then current views to explore the future. We believe that to be most useful, the learner should be given access to purposes, formats, content, contexts, and timing of assessments, as well as results and follow-up. Because the Seminar discussions fortuitously overlapped with the dramatic acceleration of advances in technology, such as artificial intelligence (AI) options, the promise of integration with assessment and learning became more feasible. The justification for our recommendations in the form of principles has many strands, but a common thread is clarifying what the assessments are about, attributes of their design, how they will be applied and scored, and how learning partners, that is, students and teachers, can benefit from their use far beyond the assessment itself. These principles overlap, with linked concepts to strengthen their effects. We now turn to the recommended principles.

Principle 1: Assessment transparency provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.

Let's start with transparency. In practice, formal and classroom assessments are often presented either as surprises or with a level of secrecy prior to their use. As a result, students may be uncertain about why they are being assessed and on what topics and skills. Our notion is that the assessment content should be obvious to the students, not something they "psych out," or feel anxious or puzzled

about. Transparent assessment practice should include sharing with the learners understandable models or guidelines for impending assessments. Transparent design means that the rules or parameters of the assessments (what's fair to be in them and what's fair to exclude) should be made clear for the boundaries of subject matter domains, as well as the types of thinking skills, assessment tasks, and criteria for evaluating the learners' performance. This clarity may suggest the enumeration of important objectives and content, or the use of graphics, such as maps or network models such as ontologies to show assessment components and their relationships. One may provide example questions and scoring guides in advance. Ideally, examples of content of interest, relevant cognitive demands, such as explanation or problem-solving, and task formats have occurred in relevant instruction. If not, efforts should be made to support learner preparation in advance of any assessment. Assessment transparency obviously aids the alignment of objectives, instruction, assessment, and effective teaching processes, a continuing requirement of coherent learning systems.

Principle 2: Assessment focus is explicit and includes purposes, outcomes, progress indicators, and processes that can be transferred to other settings, situations, and conditions.

The goal of understanding assessments is further supported by explication of the focus and purpose of the assessments in the second principle. Assessment focus refers to when in the learning process the assessment is positioned. Focus addresses whether the assessment attends to outcomes, progress in learning, or processes to be employed in the present program and then to be transferred to other situations or domains of knowledge. We identify both progress toward outcomes and processes to assist learning. Both domain-dependent and domain-independent processes are included. The purpose of assessment concerns its intended use. Is it for external reporting, grading, or consistent with our position, "assessment to support learning." The supportive use of assessment is fundamental to our principles; however, they may be applied to assessments for other purposes. In fact, it may be efficient for assessments to share more than one purpose, for instance, assessment to support ongoing learning as well as to be used in part for external reporting. Technology-supported environments can often provide procedures to aggregate responses obtained by individuals or groups during their learning activities to allow both uses.

Principle 3: Assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition.

Effective assessment design deliberately includes key student learning processes such as motivation, attention, engagement, effort, and metacognitive self-regulation. When assessments are structured to actively engage and enhance these processes (for example, by incorporating meaningful tasks, opportunities for reflection, and timely feedback), the assessment itself becomes a catalyst for learning and skill development. This principle also addresses processes included in Principle 2. Note that these processes may be domain-specific, such as close reading of literary texts or mathematics problem-solving approaches, or may be domain-independent, that is, applied to a variety of subject matters and tasks. However, domain-independent processes may have greater utility and range and apply across topics, content domains, situations, and time. For instance, processes involving metacognition (or self-analysis by learners of their own learning behaviors) can be used across the learners' lifespan of learning. Moreover, the process of learning itself can enable application of learning processes to other situations either closely or significantly varied from the conditions of learning. In psychology, this application is known as "transfer" and to be effective, the features or attributes to be applied to different situations or content need to be taught and adapted to these new tasks and settings. Types of these transferable skills are self-assessment, planning, managing attention, engagement, motivation, handling level of effort and avoiding distraction. Other transferable processes may involve interpersonal skills, for example, performing various roles in assessment that require peer collaborations.

Principle 4: Assessments model the structure of expectations and desired learning over time.

To consider further the learning and instructional interactions with assessment, Principle 4 highlights the importance of assessments to structure learners' understanding of important knowledge and skills. The format and expectations of questions and tasks on assessments can provide models for future learning. Some writers consider the practice effects of tests as a way to build student competencies. The form of questions and the requirements of answers are basic components here, although the overall content on an assessment can help students understand relationships within a topic or domain. Some assessments also give students the criteria in advance to guide evaluation. Although related to transparency, our emphasis here is about affecting learning. When the criteria are given to learners in advance of assessment, they can shape the nature and quality of their answers or more complex performances.

Principle 5: Feedback, adaptation, and other relevant instruction should be linked to assessment experiences.

Feedback, that is, direct follow-up on assessment performance, has long been an attribute of good instruction because it provides another set of strong cues about learning. Feedback varies in specificity or detail, which influences how difficult it is for students to infer the quality of their answer (and what to do about it). For instance, feedback can be given generally about the overall level of "goodness" of a response, such as an unelaborated grade on a paper. General feedback requires learners to infer (and perhaps err) about their own strengths and weaknesses. Offering a model paper or rubric for open-ended tasks adds opportunities for engagement as learners are expected to be able to identify similar and dissimilar features between their own constructions and the rubric or modeled responses. Providing elaborated explanations permits more guidance as feedback is crafted to distinguish among appropriate and somewhat lacking performance elements needed in the response. Typically, forms of feedback which give additional instruction add substantial time to the learning/assessment process. Peer feedback is another frequent way to engage all students in learning. When feedback is supplied by technology-rich systems rather than by the classroom teacher alone, it can not only support learners but also enable feedback to the instructors about their own success, allowing their review of summaries of student progress. This feedback to teachers affords them more guidance for subsequent instruction needed to support learners. Across the board, various forms of feedback have been found to be effective components of learning.

Principle 6: Assessment equity requires fairness in design of tasks and their adaptation to permit the use with respondents of different backgrounds, knowledge, and experiences.

Central to the concerns of the Gordon Seminar has been the principle of equity, a word which has many interpretations and which we hope to clarify. Equity considers attention to different experiences of learners, including their expectations for school, peer relationships, relevant knowledge, sense of belonging, and other differences in order to adjust assessment and learning to account for varied learner features. To some, equity has been understood to mean the erasure of such individual differences. That interpretation suggests that the outcomes would need to be the same for all groups of learners. In our vision of assessment and learning, we are impelled to give opportunities that connect to the diverse students we have. True equitable assessment may not result in comparable results, especially

if learners have different levels of expertise in a domain or are best served by different task formats and support. Methods to consider in adapting assessments involve offering choices of topic or context, to allow collaborative efforts to support different strengths, or to give a range of resources in the assessment to shore up prior knowledge.

An additional requirement of learning-based assessment is the effort to adapt the assessment, instruction, and learning environments in order to adjust experiences to the students' particular strengths and growth areas. We say effort to adapt because plausible adaptations may not fit targeted students. Instructional adaptation is dependent on performance and often follows feedback. Generalizations about background and culture are subject to deliberate or implicit biases. Some attempts to adapt to diversity, such as homogeneous grouping, have drawn strong opposition by groups believing that adaptation is likely to keep students operating at lower levels of expertise. There is a range of evidence about the feasibility of adapting assessments and instruction to meet diverse learner needs. Nonetheless, feasibility alone does not guarantee enactment; time, training, and materials frequently constrain what teachers can do in real time. Assessment systems should therefore provide supports that anticipate common needs and help educators avoid systematic pitfalls for each student during learning. Adaptation is often made based on the answer to the most recent assessment task and usually only attends to content and skill levels instead of the far wider swath of individual differences. Again, giving students choices among topics or projects for learning would seem to offer one path, if we assume the choices appear to be real to students (piloting is an option here) and that students know how or have been taught to make good choices. We would expect in the future more precise adaptation will occur, adjusting to multiple variables including attention, motivation, and student preferences in addition to their prior performance. Learners should also be encouraged to approach individualized goals as well. At the heart of our conception of learning is the reality of human variance on a wide range of measured and unmeasured dimensions. Our focus on assessment and subsequent adaptation is an attempt to make the opportunities fair, and to advance and avoid systematic pitfalls for each student during learning.

Principle 7:Assessment quality and validity should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility.

Although we rely on commercial purveyors of learning and assessment, it is obligatory to determine if there is evidence of the quality of the assessment and learning, before they are used widely with students. If assessments provide inaccurate or otherwise untrustworthy results, they undermine our ability to contribute to student learning and may give the students and teachers a mistaken view. Validity experts have created various models for use, some of which depend on many students making many responses, and analyses of complex statistical models. But validity does not need to depend on statistical complexities. The key point of validity is whether one can relate findings to the purpose and focus of the assessment; in particular, the evidence (e.g., student results) should closely relate to the desired inferences to be made about purposes and intended uses. Validity models may include components such as test content (sampling and distribution), response processes stimulated by the assessment and their agreement with the assessment's intent, and adequacy of scoring procedures to determine quality of performance. There are also equity elements that include the fairness of the process and content of the assessment. In assessments for learning, however, there is a tight interaction among the learning or instructional experiences, the assessment, and subsequent additional instruction, if needed. The quality of the assessment includes standard provisions, such as accuracy, representation, and framing so that learners can respond to appropriate queries. To investigate the empirical utility of learning-embedded assessments, one could imagine experimentally comparing learning sequences with and without the embedded assessments to determine the assessments' contribution to learning. Other elaborate approaches might compare all learners' response patterns with those of students deemed successful in the program or those of other experts. In all options, the evidence collected, for example, answers to questions or produced projects, closely relates to the desired inference about student learning. Validity also applies to categories into which students might fall. If the inferences from tests or assessments place students in competency groups (e.g., excellent, needs improvement, or more instruction), then the cut-scores or boundaries used to divide students into categories will also need attention. Note also that one cannot simply adapt validity models useful for one test purpose, such as accountability testing, to apply them to assessments for learning. Many efforts to date involve using both

qualitative and quantitative data to draw inferences. Sadly, validity evidence is not often available for assessments embedded in learning systems.

Because the quality of the assessment may drive the entire process of learning and certification of students, the importance of validity or other data on assessment quality cannot be overstated. Since decisions may be made regarding external learning systems and assessments for use in schools and in homes, reports of validity and quality should be made available in common language. These nontechnical reports should detail the administration condtions, and the numbers and type of students from whom the data were collected to be sure there are essential matches with students for whom the systems are intended.

Summary

In this initial chapter we have deliberately taken a quick look at principles guiding our Gordon Seminar discussions that we have culled from a vast set of options that might have been written about assessment and learning. We noted that some of our contributors to the Handbook have created or referred to other sets of principles for assessment design and use. Of key interest to many scholars is the periodic publication of the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 2014), supported by three professional organizations, the American Educational Research Association, the American Psychological Association, and the National Council for Measurement in Education. These standards have tended to be conservative in that entries in them are made when there is consensus in the field. One can see the progress of concern for equity and fairness, first more extensively treated in 1999 and expanded in 2014. With the speed of technological innovations, such as the use of AI systems to clarify transparency, generate tasks measuring complex performance, evaluate student efforts, conduct validity studies, and prepare reports, we anticipate great change. Yet, we believe that the principles we have offered remain relevant to a wide range of contexts of learning-relevant assessments, such as individual assessments, assessments of groups or teams, and capturing students' development over time. Similarly, we believe the principles apply to assessments of content knowledge, skill development, affective and self-management processes, and zest for learning. In the subsequent chapters in these volumes, scholarly explanations and examples will be available to flesh out our initial descriptions of our principles.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment* [Technical Report]. ETS. https://www.ets.org/Media/Research/pdf/gordon_commission_technical_report.pdf

Pellegrino, J. (2014). Assessment in the service of teaching and learning: Changes in practice enabled by recommended changes in policy. *Teachers College Record, 116* (110313). https://doi.org/10.1177/016146811411601102

Selected Bibliography

Assessment in the Service of Learning

Baker, E. L., & Gordon, E. W. (2014). From the assessment of education to the assessment for education: Policy and futures. *Teachers College Record, 116,* 1–24.

Darling-Hammond, L., & Adamson, F. (2014). *Beyond the bubble test: How performance assessments support 21st-century learning.* Jossey-Bass.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice, 39*(3), 72–78.

The Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment* [Technical Report]. https://www.ets.org/Media/Research/pdf/gordon_commission_technical_report.pdf

Pellegrino, J. (2014). Assessment in the service of teaching and learning: Changes in practice enabled by recommended changes in policy. *Teachers College Record, 116*(110313). https://doi.org/10.1177/016146811411601102

Ruiz-Primo, M. A., & Furtak, E. M. (2024). Classroom activity systems to supportambitious teaching and assessment. In S. F. Marion, J. W. Pellegrino, & A. I. Berman(Eds.), *Reimagining balanced assessment systems* (pp. 93–131). National Academy of Education.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Principle 1: Assessment transparency provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.

Chung, G. K. W. K., Delacruz, G. C., Dionne, G. B., & Bewley, W. L. (2003). Linking assessment and instruction using ontologies. *Proceedings of the I/ITSEC, 25,* 1811–1822.

Clancey, W. J., & Shortliffe, E. H. (Eds.). (1984). *Readings in medical artificial intelligence: The first decade.* Addison Wesley. https://impact.dbmi.columbia.edu/~ehs7001/Clancey-Shortliffe-1984/Readings%20Book.htm

Gagné, R. M., & Briggs, L. J. (1974). *Principles of instructional design.* Holt, Rinehart & Winston.

Iseli, M. R., & Jha, R. (2016). Computational issues in modeling user behavior in serious games. In H. F. O'Neil, E. L. Baker, & R. S. Perez (Eds.), *Using games and simulations for teaching and assessment: Key issues* (pp. 21–40). Routledge.

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation in Higher Education, 39*(7), 840–852. https://doi.org/10.1080/02602938.2013.875117

Moss, C. M., & Brookhart, S. M. (2012). *Learning targets: Helping students aim for understanding in today's lesson.* ASCD.

Principle 2: Assessment focus is explicit and includes purposes, outcomes, progress indicators, and processes that can be transferred to other settings, situations, and conditions.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Complete edition). Longman.

Andrade, H. L., Bennett, R. E., & Cizek, G. J. (Eds.). (2019). *Handbook of formative assessment in the disciplines* (1st ed.). Routledge. https://doi.org/10.4324/9781315166933

Armour-Thomas, E., & Gordon, E. W. (2025). *Principles of dynamic pedagogy: An integrative model of curriculum instruction and assessment for prospective and in-service teachers.* Routledge.

Chatterji, M. (2025). *User-centered assessment design: An integrated methodology for diverse populations.* Guilford Press.

Heritage, M. (2021). *Formative assessment: Making it happen in the classroom* (2nd ed.). Corwin.

Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *The Journal of Negro Education, 67*(3), 268–279. https://www.jstor.org/stable/2668195?origin=crossref

van Merriënboer, J. J. G., & Kirschner, P. A. (2007). *Ten steps to complex learning: A systematic approach to four-component instructional design.* Routledge.

Principle 3: Assessment design supports learners' processes, such as motivation, attention,engagement, effort, and metacognition.

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*(3), 261–271.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). National Academy Press.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438–481.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066x.34.10.906

Plass, J. L., & Kalyuga, S. (2019). Four ways of considering emotion in Cognitive Load Theory. *Educational Psychology Review, 31*(2), 339–359. https://doi.org/10.1007/s10648-019-09473-5

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Principle 4: Assessments model the structure of expectations and desired learning over time.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Black, P., Wilson, M., & Yao, S.-Y. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives, 9*(2–3), 71–123.

Darling-Hammond, L., Herman, J., Pellegrino, J. W., Abedi, J., Aber, J. L., Baker, E., Bennett, R., Gordon, E., Haertel, E., Hakuta, K., Ho, A., Linn, R. L., Pearson, P. D., Popham, W. J., Resnick, L., Schoenfeld, A. H., Shavelson, R., Shepard, L. A., Shulman, L., & Steele, C. M. (2013). *Criteria for high-quality assessment.* Stanford Center for Opportunity Policy in Education.

Gordon, E. G., & Bridgall B. L. (Eds.). (2006). *Affirmative development: Cultivating academic ability, critical issues in contemporary American education series*. Rowman & Littlefield Publishers, Inc.

Leonard, W. H., & Lowery, L. F. (1984). The effects of question types in textual reading upon retention of biology concepts. *Journal of Research in Science Teaching, 21*(4), 377–384. https://doi.org/10.1002/tea.3660210405

Phelps, R. P. (2012). The effects of testing on student achievement, 1910–2010. *International Journal of Testing, 12,* 21–43.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Principle 5: Feedback, adaptation, and other relevant instruction should be linked to assessment experiences.

Hattie, J. (2023). *Visible learning: The sequel: A synthesis of over 2,100 meta-analyses relating to achievement* (1st ed.). Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79–97. https://doi.org/10.3102/00346543058001079

Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20*(2), 179–189.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10,* Article 3087. https://doi.org/10.3389/fpsyg.2019.03087

Principle 6: Assessment equity requires fairness in design of tasks and their adaptation to permit the use with respondents of different backgrounds, knowledge, and experiences.

Armour-Thomas, E., McCallister, C., Boykin, A. W., & Gordon, E. W. (Eds.). (2019). *Human variance and assessment for learning.* Third World Press.

Bennett, R. E. (2023). *T*oward a theory of socioculturally responsive assessment. *Educational Assessment, 28*(2), 83–104. https://doi.org/10.1080/10627197.2023.2202312

Duran, R. P. (1989). Testing of linguistic minorities. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 573–587). Macmillan.

Gordon, E. W. (1995). Toward an equitable system of educational assessment. *The Journal of Negro Education, 64*(3), 360–372.

Herman, J. L., Bailey, A. L., & Martinez, J. F. (2023). Introduction to the special issue: Fairness in educational assessment and the next edition of the standards. *Educational Assessment, 28*(2), 65–67. https://doi.org/10.1080/10627197.2023.2215979

Nasir, N. S., Lee, C. D., Pea, R., & McKinney de Royston, M. (Eds.). (2020). *Handbook of the cultural foundations of learning.* Routledge.

Oakes, J. (1986). Keeping track, part 1: The policy and practice of curriculum inequality. *Phi Delta Kappan, 68*(1), 12–17. https://www.jstor.org/stable/20403250

Shepard, L. A. (2021). Ambitious teaching and equitable assessment: A vision for prioritizing learning, not testing. *American Educator, 45*(3), 28–37, 48.

Sireci, S. G., & Randall, J. (2021). *Evolving notions of fairness in testing in the United States in the history of educational measurement.* Routledge.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32*(2), 3–13. https://doi.org/10.3102/0013189x032002003

Principle 7: Quality and validity should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Linn, R. L. (2010). Validity. In B. McGaw, P. L. Peterson, & E. L. Baker (Eds.), *International encyclopedia of education* (3rd ed., Vol. 4, pp. 181–185). Elsevier.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

Mislevy, R. J., Oliveri, M. E., Slomp, D., Crop Eared Wolf, A., & Elliot, N. (2025). An evidentiary-reasoning lens for socioculturally responsive assessment. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 199–241). Routledge/Taylor & Francis.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–67.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement, 50*(1), 99–104.

Arguments in Support of Innovating Assessments

James W. Pellegrino

Abstract

This introductory chapter establishes assessment as a process of reasoning from evidence and presents the main arguments for why we need to innovate assessments, especially if they are to serve in support of learning. The first argument is that assessment should measure what matters, not just what is easy to measure. This means expanding the range of educational outcomes we assess to include the complex cognitive, socio-cognitive, and socio-emotional constructs that are essential for success in the worlds of today and tomorrow. The second argument is that we need new assessment design approaches and methods that leverage the affordances of digital technology to provide rich, meaningful, and useful sources of data and information. Following from the first two arguments, the third is that assessments should measure what matters and measure it well. Careful attention must be paid to the issues of validity and comparability when complex constructs are targeted for assessment, and when new tasks and tools are used for generating and interpreting evidence about student knowledge and skills.[1]

1 The structure and substance of this paper draws heavily from a previously published chapter by the author in the volume entitled Innovating Assessments to Measure and Support Complex Skills edited by Natalie Foster and Mario Piacentini and published by OECD in 2023. The author greatly appreciates permission from the Editors of the OECD volume to adapt the previously published work for use in the present context.

This *Handbook* continues arguments about assessment innovation and use that have been discussed in recent volumes such as *Innovating Assessments to Measure and Support Complex Skills* (Foster & Piacentini, 2023); *Classroom-based Assessment in STEM: Contemporary Issues and Perspectives* (Harris et al., 2023); and *Reimagining Balanced Assessment Systems* (Marion, Pellegrino, & Berman, 2024). The three volumes of this *Handbook* further contribute to those prior discussions by their collective attempt to tackle and broaden: 1) the "what" of assessment; 2) the "how" of assessment; and/or 3) the "value proposition" of assessment, i.e., the interpretation and use of results from innovative assessments but with a particular focus on for whom we measure and the interpretive value of the information obtained therein.

To develop and elaborate the three main arguments of the *What*, the *How*, and the *Value Proposition* we begin with a brief discussion of a fundamental conception about assessment, namely that it constitutes a process of reasoning from evidence guided by theory and research on critical aspects of the acquisition and development of knowledge and skill. This fundamental principle provides a basis for developing each of the three arguments noted above, including their elaboration in the multiple chapters in the three volumes of this *Handbook*. We conclude this chapter with an additional argument of consequence for educational policy and practice—to achieve innovation in assessment and effect positive impact on educational outcomes, more coherent systems of assessment are needed. Such systems better connect assessments to one another given their intended interpretive uses regarding the constructs that matter and their relationship to curriculum and instruction, respectively.

Assessment as a Process of Reasoning from Evidence

Educators assess students to learn about what they know and can do, but assessments do not offer a direct pipeline into a student's mind. Assessing educational outcomes is not as straightforward as measuring height or weight; the attributes to be measured are mental representations and processes that are not outwardly visible. Thus, an assessment is a tool designed to observe students' behaviour and produce data that can be used to draw reasonable inferences about what students know. Deciding what to assess and how to do so is not as simple as it might appear.

The process of collecting evidence to support inferences about what students know and can do represents a chain of reasoning from evidence about student competence that characterises all assessments from classroom quizzes and standardised tests to computerised tutoring programmes, to the conversation a student has with her teacher as they work through a math problem or discuss the meaning of a text. The first question in the assessment reasoning process is: "evidence about what?" *Data* become *evidence* in an analytic problem only when one has established their relevance to a conjecture being considered (Schum, 1987). Data do not provide their own meaning; their value as evidence can arise only through some interpretational framework. What a person perceives visually, for example, depends not only on the data she receives as photons of light striking her retinas, but also on what she thinks she might see. In the present context, educational assessments provide data such as written essays, marks on answer sheets, presentations of projects, or students' explanations of their problem solutions. These data become evidence only with respect to conjectures about how students acquire knowledge and skill.

In the *Knowing What Students Know* report (Pellegrino et al., 2001), the process of reasoning from evidence was portrayed as a triad of three interconnected elements: the assessment triangle. The vertices of the *assessment triangle* represent the three key elements underlying any assessment (see Figure 1): a model of student *cognition* and learning in the domain of the assessment; a set of assumptions and principles about the kinds of *observations* that will provide evidence of students' competencies; and an *interpretation* process for making sense of the evidence considering the assessment purpose and student understanding. These three elements may be explicit or implicit, but an assessment cannot be designed and implemented, or evaluated, without consideration of each. The three are represented as vertices of a triangle because each is connected to and dependent on the other two. A major tenet of the *Knowing What Students Know* report is that for an assessment to be effective and valid, the three elements must be in synchrony. The assessment triangle provides a useful framework for analysing the underpinnings of current assessments to determine how well they accomplish the goals we have in mind, as well as for designing future assessments and establishing their validity (e.g., see Marion and Pellegrino, 2006; Pellegrino et al., 2016).

The *cognition* corner of the triangle refers to theory, data, and a set of assumptions about how students represent knowledge and develop competence in an intellectual domain (e.g., fractions, Newton's laws, or thermodynamics) or regarding a socio-emotional skill or capacity. In any particular assessment application, a theory of competence in the domain of assessment is needed to identify the set of knowledge and skills that is important to measure for the intended context of use, whether that be to characterise the competencies students have acquired at some point in time to make a summative judgment, or to make formative judgments to guide subsequent instruction so as to maximise future learning. A central premise is that the theory should represent the most scientifically credible understanding of typical ways in which learners represent knowledge and develop expertise in a domain.

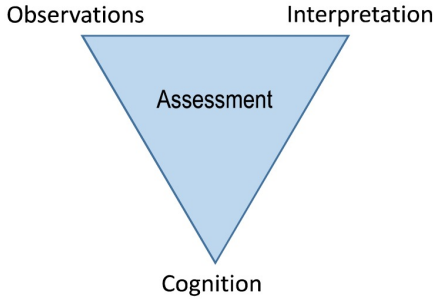
Figure 1

The Assessment Triangle (adapted from Pellegrino et al., 2001)

Cognition
Theories, models & data about how students represent knowledge & develop competence in a domain of instruction and learning.

Observations
Tasks or situations that allow one to observe students' performance.

Interpretation
Methods for making sense of the evidence coming from students' performances.



Every assessment is also based on a set of assumptions and principles about the kinds of tasks or situations that will prompt students to say, do, or create something that demonstrates important knowledge and skills. The tasks to which students are asked to respond on an assessment are not arbitrary; they must be carefully designed to provide evidence that is linked to the cognitive model of learning and to support the kinds of inferences and decisions that will be made based on the assessment results. The observation vertex of the assessment triangle represents a description or set of specifications for assessment tasks that will elicit illuminating responses from students. In assessment, one has the opportunity to structure some small corner of the world to make observations.

The assessment designer can use this capability to maximise the value of the data collected, as seen through the lens of the underlying assumptions about how students learn in the domain.

Every assessment is also based on certain assumptions and models for interpreting the evidence collected from observations. The *interpretation* vertex of the triangle encompasses all the methods and tools used to reason from fallible observations. It expresses how the observations derived from a set of assessment tasks constitute evidence about the knowledge and skills being assessed. In the context of large-scale assessment, the interpretation method is usually a statistical model, which is a characterisation or summarisation of patterns one would expect to see in the data given varying levels of student competency. In the context of classroom assessment, the interpretation is often made less formally by the teacher and is often based on an intuitive or qualitative model rather than a formal statistical one. Even informally, teachers make coordinated judgments about what aspects of students' understanding and learning are relevant, how a student has performed on one or more tasks, and what the performances mean about the state of a student's knowledge and understanding.

A crucial point is that each of the three elements of the assessment triangle not only must make sense on its own but also must connect to each of the other two elements in a meaningful way to lead to an effective assessment and sound inferences. Thus, to have a valid and effective assessment, all three vertices of the triangle must work together in synchrony.

Argument 1: Measuring What Matters

Education research has well established that teachers, students and local and national policy makers take their cues about the goals for instruction and learning from the types of tasks found on state, national, and international assessments. Thus, what we choose to assess in areas such as science, mathematics, literacy, history, problem solving, collaboration, and critical thinking is what will end up being the focus of instruction. It is therefore critical that our assessments best represent the forms of knowledge and competency and the kinds of learning we want to emphasise in our classrooms if students are to achieve the complex, multi-dimensional proficiencies needed for the worlds of today and tomorrow. Doing so, however, requires that we move away from *measuring what is easy to measuring what matters*.

There is an increasing push to encourage students to develop "21st-century skills" that combine habits of mind and that include social and affective competencies (e.g., Bellanca, 2014; Pellegrino and Hilton, 2012). The European Commission's *Rethinking Education* (2012) reform effort emphasizes the need to promote transversal skills in education, such as critical thinking and problem solving. Additionally, PISA–the international assessment of student abilities administered by the OECD–has begun testing broader competencies that go beyond the disciplinary areas of mathematics, reading and science, such as problem solving and collaborative problem solving. Such 21st-century skills–or 21st-century competencies–are deemed necessary to prepare a global workforce to succeed in a new information-driven economy. Individuals must have the problem-solving, critical thinking, and collaboration and communication skills to evaluate and make sense of new information and to act upon this information in a range of settings.

Business leaders, educational organisations and researchers have begun to call for new education policies that target the development of such broad, transferable skills and knowledge. For example, the US-based Partnership for 21st-Century Skills (2010) argues that student success in college and careers requires four essential skills: critical thinking and problem solving, communication, collaboration, and creativity and innovation. The NRC Report *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st-Century* (Pellegrino and Hilton, 2012) argued that the various sets of terms associated with the "21st-century skills" label reflect important dimensions of human competence that have been valuable for many centuries, rather than skills that are suddenly new, unique, and valuable today. The important difference across time may lie in society's desire for all students to attain levels of mastery–across multiple areas of skill and knowledge–that were previously unnecessary for individual success in education and the workplace. At the same time, the pervasive use of new digital technologies has increased the pace of communication and information exchange throughout society with the consequence that all individuals may need to be competent in processing multiple forms of information to engage in critical thinking and accomplish tasks that may be distributed across contexts that include home, school, the workplace and social networks (see e.g., Zlatkin-Troitschanskaia, Pellegrino, & Blatnik, in press).

To shift from policy into practice, assessments need to be able to measure these skills and competencies. To do that we need to have clear conceptions and definitions of the constructs to be assessed (the Cognition), the forms of evidence associated with those constructs (the Observations), and ways to make sense of that evidence for the purposes of reporting and use (the Interpretation).

Many of the *Handbook's* chapters explicitly focus on the "what" of educational assessment–the key constructs that we should be interested in assessing, why those constructs are important, and where we stand with respect to assessing them given the current educational assessment landscape. The bulk of the argument is that we should be focused on complex cognitive and socio-cognitive constructs, both within and across-disciplinary domains. The chapters discuss what we mean by these constructs and the types of tasks and situations where individuals would be required to exercise the requisite competencies, thereby providing the types of evidence that would be valid, interpretable and useful whether the intended use is at the classroom level to guide learning and instruction or in a large-scale educational monitoring context. Several chapters illuminate ways in which we might conceptualise and operationalise these constructs as well as some of the challenges in doing so. They set the stage for other chapters that move beyond conceptualisation of what we may want and need to assess as part of the advancement of 21st-century education, to the details of the design process and ways in which technology can enable the creation of situations that will provide the evidence we need while also assisting in the process of making sense of that evidence.

Argument 2: Assessment Design Processes and Applications of Technology

While it is especially useful to conceptualise assessment as a process of reasoning from evidence, the design of an actual assessment is a challenging endeavour that needs to be guided by theory and research about cognition as well as practical prescriptions regarding the processes that lead to a productive and potentially valid assessment for a particular context of use. As in any design activity, scientific knowledge provides direction and constrains the set of possibilities, but it does not prescribe the exact nature of the design, nor does it preclude ingenuity to achieve a final product. Design is always a complex process that applies theory and research to achieve near-optimal solutions under a series of multiple constraints, some of which are outside the realm of science. In the case of educational assessment, the design is influenced in important ways by variables such as its purpose (e.g., to assist learning, to measure individual attainment, or to evaluate a programme),

the context in which it will be used (e.g., classroom or large-scale), and practical constraints (e.g., resources and time).

Recognising that assessment is an evidentiary reasoning process, it has proven useful to be more systematic in framing the process of assessment design as an Evidence-Centered Design process (e.g., Mislevy & Haertel, 2006; Mislevy and Riconscente, 2006). The process starts by defining the claims that one wants to be able to make about student knowledge and the ways in which students are supposed to know and understand some particular aspect of a content domain. Examples might include aspects of algebraic thinking, ratio and proportion, force and motion, heat and temperature, etc. The most critical aspect of defining the claims one wants to make for purposes of assessment is to be as precise as possible about the elements that matter and express these in the form of verbs of cognition that are much more precise and less vague than high-level cognitive, superordinate verbs such as know and understand. Example verbs might include compare, describe, analyse, compute, elaborate, explain, predict, justify, etc. Guiding this process of specifying the claims is theory and research on the nature of domain-specific knowing and learning.

While the claims one wishes to make or verify are about the student, they are linked to the forms of evidence that would provide support for those claims–the warrants in support of each claim. The evidence statements associated with given sets of claims capture the features of work products or performances that would give substance to the claims. This includes which features need to be present and how they are weighted in any evidentiary scheme, i.e., what matters most and what matters least, or not at all. For example, if the evidence in support of a claim about a student's knowledge of the laws of motion is that the student can analyse a physical situation in terms of the forces acting on all the bodies, then the evidence might be a free body diagram that is drawn with all the forces labelled including their magnitudes and directions.

The precision that comes from elaborating the claims and evidence statements associated with a domain of knowledge and skill pays off when one turns to the design of tasks or situations that can provide the requisite evidence. In essence, tasks are not designed or selected until it is clear what forms of evidence are needed to support the range of claims associated with a given assessment situation. The tasks need to provide all the necessary evidence and they should

allow students to "show what they know" in ways that are as unambiguous as possible with respect to what the task performance implies about student knowledge and skill, i.e., the inferences about student cognition that are permissible and sustainable from a given set of assessment tasks or items.

In the *Knowing What Students Know* report (Pellegrino et al., 2001), many of the affordances of technology for advancing assessment design and practice were discussed in terms of the three interconnected components of the assessment triangle. The brief discussion that follows focuses on the constructs that could be represented in innovative assessment frameworks (Cognition), the ways in which those constructs could be realised in the assessment environment (Observation), and some of the interpretive challenges and solutions associated with doing so for purposes of measurement and reporting (Interpretation).

The Cognition vertex of the assessment triangle

What matters in assessment is what we are trying to reason about–the contemporary conception of student Cognition in a domain that matters to domain experts, educators and society. As the conception of student cognition changes and expands in terms of what students are supposed to know and be able to do, as has been the case for many domains, technology affords opportunities for substantially changing and extending the *Observation and Interpretation* components of the assessment triangle to more adequately represent and provide evidence about the constructs of interest. Doing so enhances the entire evidentiary reasoning process and the validity of an assessment given its intended interpretive use.

The Observation vertex of the assessment triangle

Technology provides opportunities for the presentation of dynamic stimuli (e.g., videos, graphics, 2- and 3-D simulations) that can be interacted with in the service of eliciting relevant sets of responses from students. Simultaneously, technology enables the generation and capture of a variety of response products, including situations in which students generate responses using multiple modalities (e.g., drawing and writing). Technology-enhanced assessments enable engagement with a variety of content and practices by opening the door to interactive stimulus environments and response formats that better match the intended reasoning and response processes that form the basis for desired claims about student proficiency (Gorin and Mislevy, 2013).

Students' interactions with these technology-enhanced assessments can be logged to provide data on how they engage in particular processes. For various 21st-century competencies, the process by which one completes the activity can be as important a piece of information about knowledge and skill as the final product. In these cases, understanding the operations that students performed in the process of creating the final product may be critical to evaluating students' proficiency. Log data offer the opportunity to reveal these actions, including where and how students spend their time, and what choices they make in situations like using a simulation. Such applications offer the potential to provide large volumes of "clickstream" and other forms of response process data that might be useful for making inferences about student thinking (Ercikan and Pellegrino, 2017).

The Interpretation vertex of the assessment triangle

Technology offers significant opportunities to enhance the reasoning-from-evidence process given the types of observations described above. Collecting these types of data makes little sense unless there are ways to reliably and meaningfully interpret them. This can evolve through mechanisms such as automated scoring of responses and application of complex parsing, statistical and inferential models for response process data (see Ercikan and Pellegrino, 2017). Critical data to consider include the time taken to perform various actions, the actual activities chosen, and their sequence and organisation. The potential exists for examining the global and local strategies students use while solving assessment problems and their implications, including how such strategies relate to the accuracy or appropriateness of final responses. Although capturing such data in a digital environment is "easy," making sense of the data is far more complicated. The same can be said for capturing data to constructed response questions where students may be expressing in written and/or graphical form an argument or explanation about some social, economic or scientific problem or phenomenon, describing the design of an investigation, or representing a model of some structure or process.

The data capture contexts described above are challenging regarding scoring and interpretation. It is here that artificial intelligence and machine learning may play a significant role in future innovative assessments (see e.g., Zhai et al., 2020a,b). Developments in machine learning also may allow researchers to analyze complex response process data, including to reveal patterns that provide important insights into students' cognitive processes in problem solving (Zhai et al., 2020a, 2020b,

2021a, 2021b; Zhai, 2021). Such data may prove to be especially informative about student thinking and reasoning and thus add greatly to the knowledge gained about student competence from large-scale assessments like PISA. An interesting example was provided in a recent report by Pohl et al. (2021) who showed that differences in student response processes, when combined with scoring methods, can significantly change the interpretation of a country's performance in PISA.

In summary, digital technologies hold great promise for helping to bring about changes in assessment that many believe are necessary. Technologies available today and innovations on the immediate horizon can be used to access information, create simulations and scenarios, allow students to engage in learning games and other activities, and enable collaboration among students. Such activities make it possible to observe, document and assess students' work as they are engaged in natural activities–perhaps reducing the need to separate formal, external assessments from learning in the moment (e.g., Behrens, DiCerbo, and Foltz, 2019). Technologies will certainly make possible the greater use of formative assessment that in turn has been shown to significantly impact student achievement. Digital activities may also provide information about abilities such as persistence, creativity and teamwork that current testing approaches cannot. Juxtaposed with this promise is the need for considerable work to be done on issues of scoring and interpretation of evidence before such embedded assessment can be useful for these varied purposes. Suffice it to say that the technology and assessment field is advancing at a very rapid rate and providing potential solutions to many of the concerns and possibilities noted above. Advances in assessment are increasingly being influenced by the rapid advances in artificial intelligence and data analytics (see e.g., multiple chapters in the volumes edited by Foster & Piacentini, 2023 and by Zhai & Krajcik, 2024; as well as Zhai & Wiebe, 2023).

Developing assessments of complex cognitive competencies requires being explicit about all three elements of the assessment triangle and their inter-relationships. Multiple chapters in the Handbook address various aspects of Argument 2 regarding the observation and interpretation elements of the assessment triangle, with an emphasis on how technology can be exploited through and within a principled design process to create assessments of the complex cognitive and socio-cognitive performances that matter. Through a combination of argument and specific examples, these chapters provide support for the claim that next-generation assessments are possible but can only be generated through a highly

principled design process that makes explicit the evidentiary chain of reasoning at the core of valid assessment. The chapters also reveal the complexities that accrue in designing such assessments and then making sense of the multiple forms of evidence they can produce.

Argument 3: Valid Interpretation and Use of Results

The joint AERA/APA/NCME Standards (1999, 2014) frame validity largely in terms of "the concept or characteristic that a test is designed to measure" (1999:5). In Messick's construct-centered view of validity, the theoretical construct the test score is purported to represent is the foundation for interpreting the validity of any given assessment (Messick, 1994). For Messick, validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (1989:13). Important work has been done to refine and advance views of validity in educational measurement (see, for example, Haertel and Lorie 2004; Kane 1992, 2001, 2006, 2013; Mislevy, Steinberg and Almond, 2003). Contemporary perspectives call for an interpretive validity argument that "specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances" (Kane, 2006:23).

Kane (2006) and others (Haertel and Lorie, 2004; Mislevy et al., 2003) distinguish between: 1) the interpretive argument, i.e., the propositions that underpin test score interpretation; and 2) the evidence and arguments that provide the necessary warrants for the propositions or claims of the interpretive argument. In essence this view identifies as the two essential components of a validity argument the claims being made about the focus of an assessment and how the results can be used (interpretive argument), together with the evidence and arguments in support of those claims. Appropriating this approach, contemporary educational measurement theorists have framed test validity as a reasoned argument backed by evidence (e.g., Kane, 2006). An argument and evidence framing of validity supports investigations for a broad scope of assessment designs and purposes, including many that go beyond typical large-scale tests of academic achievement or aptitude and move one into the arena of innovative and instructionally supportive assessments (e.g., see Pellegrino et al., 2016).

Given the nature of the constructs of interest, including their inherent complexity and multi-dimensionality, we must acknowledge from the outset the challenges that will be faced in establishing validity arguments for innovative assessments of 21st-century competencies, including the reporting of results for various intended use cases. Validity arguments will depend on well-developed interpretive arguments that include: 1) clear specifications of the constructs of interest and their associated conceptual backing; 2) the forms of evidence associated with those constructs; and 3) the methods for interpretation and reporting of that evidence. Such interpretive arguments are essential to guide assessment design processes, including carefully thought-out applications of technology and data analytics to support the observational and inferential aspects of the overall reasoning from the evidence process. As noted above, carefully developed and articulated claims about what is being assessed and reported then need to be supported by empirical evidence. Such evidence can be derived from multiple forms of data involving variations in human performance and are essential to establishing an assessment's validity argument.

In pursuing innovative assessment of 21st century competencies, of paramount concern are issues of equity and fairness as part of the validity argument. Of particular concern is comparability of results and validity of inferences derived from performance obtained across different modes of assessment, especially for varying groups of students (see Berman et al., 2020). As assessment has moved from paper-and-pencil formats to digitally-based assessment, the general focus has been on mode comparability and concerns about student familiarity and differential access to the hardware and software used (see Way and Strain-Seymour, 2021). However, as the digital assessment world advances, a significant issue for innovative assessment is determining how student background characteristics including language, culture, and educational experience influence performance on different types of tasks and innovative assessment designs that leverage the power of technology. As the assessment environments and tasks become more innovative, equity and fairness concerns become even more important than general mode comparability effects. Thus, a key part of the validity argument for any innovative assessment will be establishing the socio-cultural boundaries related to equitable and fair interpretations and uses of the assessment results.

Many of the *Handbook's* chapters focus on critical aspects of design and development as part of establishing the validity of next-generation assessments for 21st-century competencies. More specifically, multiple chapters focus on the validity evidence that would be derived through the application of a principled design process that forces one to articulate, in varying degrees of detail, the connections between and among the cognition, observation and interpretation components of the assessment. Such evidence contributes to the assessment's overall validity argument but needs to be complemented by various forms of empirical data on how the assessment performs.

Towards More Coherent and Instructionally Supportive Systems of Assessment

No single assessment can evaluate all the forms of knowledge and skill that we value for students; nor can a single instrument meet all the goals held by parents, practitioners and policymakers. As argued below, it is important to envision a coordinated system of assessments in which different tools are used for different purposes—for example, formative and summative, or diagnostic vs. large-scale reporting. Within such systems, however, all assessments should faithfully represent the constructs of interest, and all should model good teaching and learning practice.

At least four major features define the elements of assessment systems that can fully reflect rigorous standards and support the evaluation of deeper learning (see Darling-Hammond et al. (2013) for an elaboration of the relevance, meaning and salient features of each of these criteria):

- *Assessment of higher-order cognitive skills* through most of the tasks that students encounter—in other words, tasks that tap the skills that support transferable learning, rather than emphasising only those that tap rote learning and the use of basic procedures. While there is a necessary place for basic skills and procedural knowledge, it must be balanced with attention to critical thinking and applications of knowledge to new contexts.
- *High-fidelity assessment of critical abilities*, as articulated in the standards—such as communication (speaking, reading, writing and listening in multi-media forms), collaboration, modelling, complex problem solving and research, in addition to key subject matter concepts. Tasks should measure these abilities directly as they will be used in the real world rather than through a remote proxy.

- *Use of items that are instructionally sensitive and educationally valuable*– in other words, tasks should be designed so that the underlying concepts can be taught and learned, distinguishing between students who have been well- or badly-taught rather than reflecting students' differential access to outside-of-school experiences (frequently associated with their socio-economic status or cultural context) or interpretations that mostly reflect test-taking skills. Preparing for (and sometimes engaging in) the assessments should engage students in instructionally valuable activities, and results from the tests should provide instructionally useful information.
- *Assessments that are valid, reliable, and fair* for a range of learners, such that they measure well what they purport to measure, be accurate in evaluating students' abilities and do so reliably across testing contexts and scorers. They should also be unbiased and accessible and used in ways that support positive outcomes for students and instructional quality.

A major challenge is determining the conditions and resources needed to create coherent systems of assessments that work across contexts ranging from the classroom to larger organisational units such as districts, states, countries and internationally. Regardless of their context of implementation, assessments in such systems must support the ambitious goals we have for the educational system, meet the information needs of different stakeholders, and align with the criteria above. The volume *Reimagining Balanced Assessment Systems* (Marion, Pellegrino, & Berman, 2024) provides a very powerful and comprehensive argument for such coherence with explicit principles for design and implementation across multiple levels of the educational system. In such balanced assessment systems all assessments are based on contemporary theory and research on knowing, learning and human development and all are focused on providing information that supports equitable and ambitious classroom teaching and learning.

Final Thoughts

Innovation and change are always challenging no matter the context. They have been especially challenging in education systems given long-standing and entrenched histories of educational policy and practice. Many have argued that education has changed little over the last 50–100 years in terms of how it is organised, delivered, what is taught and how it is assessed. Yes, there have been

changes in the subject matter learned, in the pedagogies employed and, most recently, in the uses of technology. Those changes have been evolutionary and not revolutionary. Not surprisingly, much the same can be argued about educational assessment regarding what we assess and how we do so, including applications of technology to the practice of assessment—evolutionary, but not revolutionary.

This *Handbook* is focused on an alternative and perhaps revolutionary vision that starts with the complex competencies that are deemed critical for citizens of the 21st-century. The Handbook's chapters provide a vision of what they are by characterising how we might create environments and situations where the competencies of interest would necessarily be expressed in addition to describing the evidence that those environments could provide about those competencies. Some might find it curious that a vision for the future of education starts with assessment rather than curriculum and instruction. One of the benefits of thinking first about the outcomes we desire from the educational system, with a particular focus on what they would look like, is that this information provides the basis for a "Backwards Design" process regarding the design of curriculum and instruction that can lead to those outcomes (Wiggins and McTighe, 2011).

As you read the chapters in the three volumes of this Handbook, we hope they help you consider the costs and benefits of innovative educational assessment. These considerations include the competencies described, the types of environments for assessing them, conceptual and operational design and implementation challenges, and the value of the information derived in terms of its utility for classroom teaching and learning and for education more broadly. We also suggest that you consider what it might take to move in the directions highlighted by this volume given the many entrenched assumptions, policies and practices that have come to dominate the educational assessment landscape.

References

(AERA/APA/NCME) American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999, 2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Behrens, J. T., DiCerbo, K. E., & Foltz, P. W. (2019). Assessment of complex performances in digital environments. *The Annals of the American Academy of Political and Social Science, 683*(1), 217–232.

Bellanca, J. (2014). *Deeper learning: Beyond 21st-century skills.* Bloomington, IN: Solution Tree Press.

Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (Eds.). (2020). *Comparability of Large-Scale Educational Assessments: Issues and Recommendations.* National Academy of Education.

Darling-Hammond, L., Herman, J., Pellegrino, J. W., et al. (2013). *Criteria for high-quality assessment. Stanford, CA: Stanford Center for Opportunity Policy in Education.*

Ercikan, K., & Pellegrino, J. W. (2017). *Validation of score meaning for the next generation of assessments: The use of response processes.* Taylor & Francis.

European Commission. (2012). *Rethinking education: Investing in skills for better socio-economic outcomes*.

Foster, N., & Piacentini, M. (Eds.). (2023). *Innovating assessments to measure and support complex skills.* Paris: OECD Publishing.

Gorin, J. S., & Mislevy, R. J. (2013). Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment [Paper presentation].

Haertel, E. H., & Lorié, W. A. (2004). *Validating standards-based test score interpretations. Measurement, 2*(2), 61–103.

Harris, C., Wiebe, E., Grover, S., & Pellegrino, J. W. (Eds.). (2023). *Classroom-based assessment in STEM: Contemporary issues and perspectives.* Community for Advancing Discovery Research in Education (CADRE). Education Development Center, Inc.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73.

Marion, S., & Pellegrino, J. W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice,* Winter 2006, 47–57.

Marion, S., Pellegrino, J. W., & Berman, A. (Eds.). (2024). *Reimagining balanced assessment systems.* National Academy of Education.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13–23.

Mislevy, R., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Erlbaum.

Mislevy, R. J., Steinberg, L., & Almond, R (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–67.

Partnership for 21st-Century Skills. (2010). *21st-century readiness for every student: A policymaker's guide.* Tucson, AZ: Author. Available: https://files.eric.ed.gov/fulltext/ED519425.pdf

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* National Academies Press.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550

Pellegrino, J. W., & Hilton, M. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st-century.* National Academies Press.

Pellegrino, J. W., Wilson, M., Koenig, J., & Beatty, A. (Eds.). (2014). *Developing assessments for the Next Generation Science Standards.* National Academies Press.

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science, 372*(6540), 338–340.

Schum, D. (1987). *Evidence and inference for the intelligence analyst.* University of America Press.

Way, D., & Strain-Seymour, E. (2021). *A framework for considering device and interface features that may affect student performance on the National Assessment of Educational Progress.* NAEP Validity Studies Panel.

Wiggins, G., & McTighe, J. (2011). *The understanding by design guide to creating high-quality units.* ASCD.

Zhai, X. (2021). Practices and theories: How can machine learning assist in innovative assessment practices in science education. *Journal of Science Education and Technology, 30*(2), 1–11.

Zhai, X., Haudek, K. C., Shi, L., Nehm, R., & Urban-Lurain, M. (2020a). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430–1459.

Zhai, X., Haudek, K. C., Wilson, C., & Stuhlsatz, M. (2021a). A framework of construct-irrelevant variance for contextualized constructed response assessment. *Frontiers in Education, 6,* 751283. https://doi.org/10.3389/feduc.2021.751283

Zhai, X., & Krajcik, J. (Eds.). (2024). *Uses of AI in STEM education.* Oxford University Press.

Zhai, X., Krajcik, J., & Pellegrino, J. (2021b). On the validity of machine learning-based next generation science assessments: A validity inferential network. *Journal of Science Education and Technology, 30*(2), 298–312.

Zhai, X., & Wiebe, E. (2023). Technology-based innovative assessment. In C. Harris, E. Wiebe, S. Grover, and J. W. Pellegrino (Eds.), *Classroom-based STEM assessment: Contemporary issues and perspectives,* (pp. 99–126). Community for Advancing Discovery Research in Education (CADRE). Boston: Education Development Center.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020b). Applying machine learning in science assessment: A systematic review. *Studies in Science Education, 56*(1), 111–151.

Zlatkin-Troitschanskaia, O., Pellegrino, J. W., & Bartnik, T. (in press). Learning to think critically. In R. E. Mayer, P. A. Alexander, & L. Fiorella (Eds.), *Handbook of research on learning and instruction.* New York: Routledge.

Innovating Assessment Design to Better Measure and Support Learning

Natalie Foster and Mario Piacentini

Abstract

Assessments should support deep learning as well as measure its outcomes. This chapter proposes five design principles–following insights from research in the learning and cognitive sciences–for designing innovative assessments to measure and support the development of complex competencies. These principles include: 1) using extended performance tasks; 2) accounting for students' prior knowledge; 3) providing opportunities for productive failure; 4) providing feedback and instructional support; and 5) designing "low floor, high ceiling" tasks. Our fundamental argument driving these design principles is that assessments should not only measure student progress towards educational goals, but also model and provide insight into students' deep learning processes. Adopting these design principles will enable educators to collect information about how well students can engage in complex thinking and problem-solving processes while reducing the current distance between assessment and learning. We illustrate how these design principles can be applied to assessments in different contexts and for different purposes, including a large-scale international assessment (the PISA 2025 Learning in the Digital World assessment) and a classroom-based formative assessment. We conclude by discussing areas for future research and reflecting on the different forms of capital (political, intellectual and financial) required for advancing this vision of assessment.

In recent years, researchers have argued for the need to innovate educational assessments to better measure and support the development of important skills[1] (e.g., Foster & Piacentini, 2023; Kyllonen & Sevak, 2024; Schwartz & Arena, 2013). Such works respond to shifts in educational discourse and policy about what is important to teach and learn–so-called 21st-century competencies. Innovation is required across the entire assessment development process: from conceptual foundations (defining the components of what are often complex constructs and the authentic contexts in which they are engaged), to design considerations (how tasks are designed and delivered to test takers), measurement issues (how to generate, interpret and accumulate useful evidence about what students know and can do), and reporting options (how to clearly communicate the results to intended audiences, be they teachers, learners, administrators or policy makers). To achieve positive impact on educational outcomes, more coherent systems of assessment are also needed, increasing the alignment between formative and summative assessment (Darling-Hammond et al., 2013; Pellegrino, 2023).

In this chapter, we set forth some key arguments in favor of innovating assessments to better measure and support students' learning. We then focus in on the design considerations that should guide the development of next-generation assessments that intend to measure and support the cultivation of complex skills. We propose five design principles that respond to key insights from research in the learning and cognitive sciences. These are: 1) use extended performance tasks; 2) account for student knowledge in task design and performance interpretation; 3) provide opportunities for productive failure; 4) provide feedback and instructional support during tasks; and 5) design "low floor, high ceiling" tasks. One of the key arguments of our approach is that assessments should not only be useful for measuring student progress towards educational goals, but they should also model and provide insight into students' deep learning processes (Piacentini et al., 2023). In this way we can collect important data on complex thinking and problem-solving processes, and at the same time, we can reduce the current distance between assessment and the learning we want to promote in classrooms. We illustrate how these design principles can be applied to assessments in different contexts

1 The structure and substance of this chapter draws from content in a previously published volume entitled *Innovating Assessments to Measure and Support Complex Skills* edited by Natalie Foster and Mario Piacentini and published by OECD in 2023

and for different purposes, including a large-scale international assessment (the Learning in the Digital World assessment, to be conducted in the 2025 cycle of the Programme for International Student Assessment) and a classroom-based formative assessment.

Arguments in support of innovating assessments

A fundamental conception about assessment is that it constitutes a process of reasoning from evidence, guided by theory about the critical aspects of knowledge and skill one is interested in measuring. This process of reasoning from evidence has been portrayed as a triad of three interconnected elements: the *assessment triangle* (Pellegrino et al., 2001). The vertices of the assessment triangle represent the three key elements underlying any coherent assessment: a model of student *cognition* and learning in the domain of the assessment; a set of assumptions and principles about the kinds of *observations* that will provide evidence of students' competencies; and an *interpretation* process for making sense of the evidence (See Figure 1).

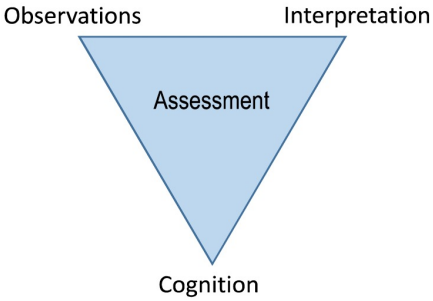
Figure 1

The assessment triangle, Source: Pellegrino et al. (2001).

Cognition
Theories, models & data about how students represent knowledge & develop competence in a domain of instruction and learning.

Observations
Tasks or situations that allow one to observe students' performance.

Interpretation
Methods for making sense of the evidence coming from students' performances.



The assessment triangle provides a useful framework for analyzing the foundations of current assessments to determine how well they accomplish the educational goals we value, as well as how we can design future assessments and establish their validity (Marion & Pellegrino, 2007; Pellegrino et al., 2016). Before we discuss how future assessments might be designed to better support learning and instruction, we first consider the extent to which current assessments are fit for

purpose given an increasing focus on preparing students for the future equipped with so-called 21st-century competencies. We also briefly discuss some of the challenges associated with developing assessments of more complex skills and competencies as well as new opportunities for innovation in test design and measurement technology.

Measuring what matters

It is no surprise that what we choose to assess ends up being the focus of instruction. By the same logic, those skills and constructs that we do not assess inevitably fall by the wayside in terms of instruction focus. Since the turn of the century, business leaders, educational organizations and researchers have increasingly called for new education policies targeting the development of broad, transferable skills and knowledge in education to prepare today's students for participation in an emerging global knowledge society. Information and communication technologies (ICTs) have radically transformed our societies, connecting people around the world, delivering unprecedented amounts of information to us, and giving rise to new forms of decentralized and autonomous learning. Young people today must not only learn to participate in a more interconnected, digital and rapidly changing world, they must also develop agency.

Several international frameworks have focused on concretely identifying the sets of knowledge, skills and attitudes–or so-called 21st-century competencies–that young people require to succeed in their futures (Binkley et al., 2011; European Commission, 2019; Fadel & Groff, 2018; OECD, 2018; Pellegrino & Hilton, 2012; Scott, 2015; World Economic Forum, 2015). Broadly speaking, some distinct categories of competencies emerge across these different frameworks including cognitive (e.g., creative thinking, critical thinking, problem solving), interpersonal (e.g., communication, collaboration), metacognitive (e.g., self-regulated learning), intrapersonal (e.g., persistence, adaptability) and digital (e.g., media literacy, digital literacy) (Foster, 2023). Digital competencies are often a central component that intersect with other 21st-century competencies given the vast proliferation of ICTs and related advances (e.g., Artificial Intelligence) in both our personal and professional lives, in turn transforming our individual and collective capacities for communication, collaboration, searching for information and using knowledge.

Yet for as much as 21st-century competencies are now a familiar component of contemporary discourse on education, assessments need to be able to measure

these competencies if we want to truly shift policy into practice. It is critical that our assessments reflect the forms of knowledge and learning we want to emphasise in our classrooms if students are to achieve the complex and multi-dimensional competencies they will need for the worlds of today and tomorrow. Not only does this mean being able to measure the extent to which students master these important competencies in different contexts, but also designing assessments that can provide timely feedback to learners and educators that can benefit instructional practices. In other words, developing assessments that *support* deep learning as well as measure its outcomes. However, as argued by Pellegrino (2023), doing so requires that we move away from *measuring what is easy* to *measuring what matters*.

Of course, measuring what matters is easier said than achieved. Several interconnected conceptual and practical challenges exist when designing assessments of 21st-century competencies, particularly in the context of large-scale assessment. These challenges include defining the assessment construct and learning progressions for complex competencies at different levels of proficiency, identifying the extent to which domain-specific knowledge supports performance and performance in an assessment can be generalized to other contexts, being able to identify and design tasks that elicit both outcomes and the processes that often define these competencies, and being able to generate interpretable evidence about students' ways of thinking and doing (see Foster, 2023, for an in-depth discussion of each). While these challenges remain real and require significant investments of time, intellectual and financial resources to overcome, some promising examples of assessments of 21st-century competencies exist at scale (see the OECD's Programme for International Student Assessment (PISA) or the Assessment and Teaching of 21st-Century Skills (ACT21S) project). What's more, advances in computer-based assessment–and particularly design and measurement technology–provide new opportunities to pursue this assessment agenda.

Unlocking new advances in design and measurement technology

Advances in technology offer much potential to transform what we can measure, how we can make sense of test taker performance, and how assessment can relate to learning (Thornton, 2012; Shute & Kim, 2013; Timmis et al., 2016; Zhai et al., 2020; DiCerbo et al., 2017). Technology can introduce new forms of active, immersive and iterative performance-based tasks in interactive environments that make it possible

to observe how test takers engage in authentic, open-ended learning and problem-solving activities. These can provide richer observations and potential evidence about students' thinking processes, behaviors and decision-making, as well as enable the measurement of dynamic skills beyond the capability of more traditional and static test items. These environments allow students to engage actively in the processes of making and doing, making it possible to track the strategies that test takers employ and the decisions they make as they work through complex tasks that evolve on their basis of their actions. Such student-centred tasks also provide opportunities for students to learn and develop skills as they make decisions and engage in iterative problem-solving. In addition to providing dynamic problems to test takers, such technology-enhanced environments can also replicate certain situations in a standardized way for assessment (e.g., simulating collaborative encounters) or visualize scenarios that would otherwise be impossible in school settings (e.g., modelling dynamic systems).

New task modalities, problem types and affordances mean there are also new possibilities for response types—and in turn, new sources of potential evidence about test takers (Sabatini et al., 2023). Digital platforms can capture, time stamp and log student interactions with the test environment. This is especially transformative in the context of measuring 21st-century competencies, as the process by which an individual engages with an activity can be just as valuable for evaluating proficiency as their final product. These process data, when coupled with appropriate analytical models, can reveal how students engage with problems, the choices they make, and the strategies they do (or do not) implement. Patterns of behavior associated with different mastery levels can be identified, which can then be used to augment the precision of performance scores and provide diagnostic information to educators about students.

Taken together, these opportunities represent a powerful shift in assessment away from simple knowledge reproduction to knowledge-in-use—exactly the types of performances in which 21st-century competencies are engaged and developed. Measuring these constructs well necessarily requires assessments to be closer tied to the processes and contexts of learning and instruction—something that technology can facilitate through embedded assessment. Embedded assessment environments can integrate pedagogical affordances like scaffolding and feedback to explicitly support learning, merging summative and formative assessment and providing measures of the capacity of test takers to learn and transfer their learning

to other tasks. By allowing for the real-time measurement of students' capacities as they engage in meaningful learning activities, technology holds the promise of creating new systems of evaluation where evidence on students' progress is collected in a continuous and unobtrusive way. In turn, educators can gain better insights into their students' learning processes, and assessment and learning are no longer explicitly separated activities.

Design principles for innovative assessments

While these opportunities are exciting, few extant assessment systems take full advantage of this potential to measure what matters. Traditional assessments (especially large-scale) have evolved to comply with practical constraints on testing time, cost efficiency and established measurement models; this often resulting in a narrow focus on knowledge mastery and response correctness, abstracting from the cognitive, interpersonal and intrapersonal processes that support test taker responses. We argue that test taker performance in the types of restricted and static situations used in traditional assessments might not provide particularly valid inferences on their capabilities to think, act and learn in real-life situations.

While the technology-enabled innovations described above clearly offer new possibilities for assessment design that can bridge this gap, it's important to acknowledge that such innovations are only useful insofar as they are integrated purposefully within a principled design process. This means that choices about what aspects of performance to simulate, what tools and affordances to include, what evidence to collect, and what interpretations to draw from the data are guided by an explicit chain of reasoning. In the rest of this chapter, we contend that it is possible to address this misalignment between current practice and promise in two ways: first, by taking stock of research that has investigated the mental structures that support the types of learning and problem solving encompassed by 21st-century competencies; and second—much more difficult—by creating an internally consistent system of teaching practices and assessments that reflect these research insights. In this system, deeper learning experiences prepare students for future learning, and innovative assessments measure how effectively students have engaged with these deeper learning experiences.

Insights about deeper learning processes and their implications for assessment

One of the main conclusions from research in the learning sciences is that learning is a socially situated process (Dumont et al., 2010; Darling-Hammond et al., 2019). People learn and develop expertise by participating in the practices of a community of experts and by learning to use the tools, languages and strategies that have been developed within that community (Mislevy, 2018; Ericsson, 2006; Pellegrino & Hilton, 2012). Becoming skillful in any domain involves learning the ways of thinking and acting that are aligned with those valued by a community of experts and by soliciting and using their feedback. That learning is mediated by socially constructed practices has clear implications for instruction: deeper learning occurs when students engage in activities that are realistic, complex, meaningful, and motivating, and when they can call upon the experience of knowledgeable others for guidance and support. Assessment experiences aimed at engaging and measuring deeper learning processes must therefore replicate these features.

Some research has focused on contrasting "experts" (i.e., individuals that have constructed mental models in a given domain and who, through participation in key practices, have learned to apply these ideas to solve new problems) with "novices" (i.e., those who have not consolidated their basic knowledge in a domain through practice). A recurrent observation is that experts have strong metacognitive skills (Hatano, 1990). While learning and problem solving, they engage in regulatory behaviors such as knowing when to apply a procedure, planning, predicting the outcomes of an action, questioning the limitations of their knowledge, monitoring their progress, and efficiently apportioning cognitive and emotional resources. The capacity to regulate one own's learning and adapt accordingly further distinguishes routine experts from adaptive experts, with the latter being "characterized by their flexible, innovative, and creative competencies within the domain" (Hatano & Oura, 2003, p. 28). Instructional and assessment practices that aim to develop and differentiate experts and novices must provide learners with opportunities to engage these regulatory behaviours.

Research has also shown that general problem-solving procedures ("weak methods"), such as trial-and-error or hill climbing, are slow and inefficient (Pellegrino et al., 2001), and that experts instead use deep knowledge of the domain ("strong methods") to solve problems. This deeper knowledge does not refer to isolated facts, but rather to knowledge encoded in a way that closely links it with

its contexts of practice and conditions of use. When experts face new problems or scenarios, they can readily activate and retrieve the subset of their knowledge that is relevant to the task at hand (Simon, 1980). Learners progress in their mastery of a discipline through similar processes of acquisition and use of increasingly well-structured knowledge schema in different contexts. The implication for assessment here is that learning and skill mastery cannot readily be divorced from their specific contexts of practice, and as such, measuring deeper learning processes must occur in context rather than in highly generalised problem contexts.

Design innovations to measure learning and support teaching

Thus far, we have described the types of competencies that students need to learn, problem-solve, and ultimately thrive in their future roles within society. Thanks to extensive research in the learning sciences, we also increasingly understand aspects of the processes involved in developing these higher order thinking and learning skills, and the types of experiences that elicit them. In this section, we highlight five broad design principles that we argue should form the basis of "next-generation assessments" that can yield potentially valid evidence about where students are in the development of 21st-century competencies (in summative applications) and about what they need to do to progress in these skills (in formative applications). For a more detailed discussion and examples of these principles, see Piacentini et al. (2023).

These five design principles also closely connect to the Principles set forth in and throughout this Handbook, namely that the focus of assessment should target and consider progress, outcomes and processes that can be transferred to other settings, situations and conditions (Principle 2), and that the design of assessments should support learners' processes, motivation, attention, engagement, effort and metacognition (self-regulation) (Principle 4).

Design Principle 1: Include Extended Performance Tasks

Assessments that aim to measure how prepared students are for deeper learning–and to generate insights useful to teaching and learning processes–must engage students in active and authentic learning processes. Students engage 21st-century competencies in situations in which they interact with others, evaluate available resources, make choices about the course of action to take, try out strategies, iterate, and adapt according to the results. From an assessment perspective,

this means providing students with a purposeful challenge that replicates the key features of those educational experiences where deeper learning happens. Evaluating students' capacity to construct new knowledge in choice-rich environments means that students should be given sufficient time and affordances to demonstrate what they can do. We argue that extended units with multiple activities, sequenced as steps towards achieving a main learning goal, can provide a more authentic and motivating experience of assessment that provides valid data about students' competencies (i.e., evidence that is predictive of what students can do outside of the constrained and stressful context of a test).

Advances in technology now allow much more data to be captured on how students spend their time in extended tasks by immersing them in simulated environments and communities of practice. These environments can facilitate a more open interpretation of students' goals and their exploration of the problem constraints, reward diverse solution strategies and outcomes, and provide feedback to learners. This also makes it possible to observe metacognitive processes that are crucial to learning in a non-obtrusive way, tracking how students plan and implement strategies, how they behave when they are stuck, and how they respond to feedback (Nunes et al., 2003). The application of a principled design process can lead to a productive use of these process data to augment the evidence that is derived from final solutions, therefore reducing the trade-off between reliability and authenticity (e.g., Piacentini, 2023; Sabatini et al., 2023).

Design Principle 2: Account for Knowledge in Task Design and Reporting of Performance

Knowledge plays an important role in authentic task performance. Competencies like creativity, critical thinking or communication are rarely exercised within a vacuum. In an assessment context, students' ability to demonstrate these skills will always be observed within a given context and their knowledge about this context or situation will influence the type of strategies they use as well as what they are able to accomplish (Mislevy, 2018). Attempting to design completely decontextualized assessment problems or scenarios threatens validity: if no knowledge is required to solve a task, can an assessment really measure the types of complex competencies it claims to be interested in?

What this means, in turn, is that it is important to explicitly identify the knowledge that students need in any assessment context to meaningfully engage with the

test activities and to evaluate the extent to which differences in prior knowledge influence the evidence we can obtain on the target constructs. It is also important to assess complex skills across a variety of knowledge and application domains to make valid conclusions on students' mastery of these skills. Evaluating the extent to which students possess relevant knowledge when engaging with a complex performance task (for example, through a short battery of items) should become an integral part of the design and assessment process in next-generation assessments, as this can help to interpret their subsequent behaviors and choices during the assessment (Piacentini, 2023; Roll & Barhak-Mirkowitz, 2023).

Design Principle 3: Provide Opportunities for Productive Failure

There is evidence that we can make robust claims about students' preparedness to learn new things by studying how they work on unfamiliar problems (Roll et al., 2011; Schwartz & Martin, 2004). For example, "invention activities" ask students to work on problems requiring concepts or procedures that they have not yet been taught, with the aim that students explore and understand the core properties of a construct before being taught expert solutions and strategies (Roll et al., 2012). Students often fail in their attempts to solve or generate canonical solutions to such problems, but experimental evidence shows that students who learn through invention activities are better at transferring their knowledge (i.e., solving other tasks requiring the same knowledge schemes in a different context) in comparison to students who are directly told how to solve the problem (Loibl et al., 2016; Kapur & Bielaczyc, 2012). Invention activities therefore help students to deeply understand concepts, identify the limitations of previous procedures when they do not work, and look for new patterns and interpretations that connect with and build upon their existing knowledge. Similar ideas can also be found in other "teaching for understanding" frameworks (e.g Wiske, 1997). In an assessment context, these types of activities can provide evidence about whether students can flexibly apply their knowledge schema to unfamiliar contexts.

In traditional tests, if students do not know the relevant procedure to follow there is little they can do to progress (Schwartz & Arena, 2013). Yet in the real world, we can access resources and ask more knowledgeable others for help when learning and problem solving. Assessments that challenge students to learn to solve new problems should incorporate resources for learning, because problem solving always requires some degree of knowledge. These should be carefully

crafted so that they do not provide prescriptive solution steps, but rather provide opportunities to learn about core properties of the problem and encourage implementing a certain strategy that helps to make progress toward a solution. Contrasting cases represent one approach to providing such structure in learning and invention activities that has proven effective in experimental settings and that could be applied to larger-scale assessments (e.g., Shwartz & Martin, 2004; Taylor et al., 2010).

Design Principle 4: Provide Feedback and Instructional Support to Students During Tasks

To complement the design principle 3, instructional support in the form of advice, feedback, or prompts as students engage in activities can promote deep learning in beginners and enable the decisions they make in their learning to be observed (Azevedo & Aleven, 2013; van Joolingen et al., 2005). Targeted feedback and scaffolded interventions can also reduce the risk that beginners disengage from an assessment because they perceive it to be beyond their capacities: it can provide clarity over task instructions, reduce the degrees of freedom or number of acts required to make progress, signal critical features that a student may have missed upon first attempt, reproduce partial solutions, and elicit further articulation or reflection questions (Guzdial et al., 2001). All these are important functions that can serve to elicit more information about students' competency level and reengage students if they appear disengaged–which is especially important in the context of extended performance tasks with less discrete items and fewer clear "data points" to inform scores.

Including feedback and instructional support in summative tests may be challenged because of fairness (e.g., unfairly penalizing students who do not need such supports) or validity concerns. However, Shute et al. (2008) evaluated the psychometric quality of an algebra assessment delivered by a digital learning system that combined adaptive task sequencing with instructional feedback: a comparison of metrics for a treated (with feedback) and control (without feedback) sample showed that providing instructional support did not make the assessment less able to detect differences between students. From a measurement perspective, the main challenge remains to develop and validate psychometric models that account for how these resources affect students' measured ability, as this ability is no longer fixed but can progress during the assessment experience (see Piacentini, 2023, and Scalise et al., 2023, for a more in-depth discussion of this issue).

Design Principle 5: Design Tasks That Are 'Low Floor, High Ceiling' or Otherwise Adaptive

Next-generation assessments of 21st-century competencies should enable all students to demonstrate their ability to learn and tackle problems by using the tools and resources available to them—regardless of their initial level of knowledge or skill. Adapting assessment challenges to different abilities not only improves the quality of the measures but also the authenticity of the assessment experience: in real life, people seldom take on challenges that are too easy or impossible to achieve, yet in traditional tests this happens quite frequently. One approach to catering to different student ability levels is to design so-called "low floor, high ceiling" tasks, meaning that they are largely accessible to all ability levels while providing scope to challenge top performers. One cluster of low floor, high ceiling problems asks students to produce an original artifact: for example, a story, a game, a design for a new product, an investigative report, a speech, etc. These artifacts generate a wide range of qualitatively distinct responses and even top performers are incentivized to produce a solution that is richer, more complete, and unique. This type of design can also be used in more standardized problem-solving tasks if students are informed about intermediate targets to achieve and that progress towards more sophisticated solutions will be rewarded.

Adaptive designs can also address the complexity of measuring learning-in-action amongst heterogeneous student groups. A relatively simple way to integrate some level of adaptivity in assessment design is to structure achieving larger and more complex goals as a sequence of tasks that gradually increase in difficulty (similar to a "level-up" mechanism) and instruct students to complete as many as they can. More proficient students will quickly complete the initial set of simple tasks, after which they will encounter problems that challenge them; less prepared students are still able to engage meaningfully with the tasks and demonstrate what they can do, even if they do not complete the full sequence. Both groups of students work at the cutting edge of their abilities with obvious benefits in terms of measurement quality and test engagement. More sophisticated adaptive pathways could also be introduced: based on the quality of students' work in previous tasks, they could be directed on-the-fly towards easier or more difficult subsequent tasks.

From principles to practice

These five design principles are intended as broad guidelines for designing next-generation assessments capable of measuring and supporting the development of 21st-century competencies. They are not intended to be prescriptive but rather illustrate the characteristics of innovative assessments that can provide valid information about students' real-life abilities. Designing assessments is a complex and challenging endeavor that must be guided by theory and research about cognition; yet just like any other design activity, scientific knowledge provides direction and constrains the set of possibilities, but it does not prescribe the exact nature of the design nor ingenuity in the final product. In the case of educational assessment, the design is influenced in important ways by variables such as its purpose (e.g., to assist learning, to measure individual attainment, or to evaluate a program), the context in which it will be used (e.g., classroom or large-scale), and practical constraints (e.g., resources and time) (Pellegrino, 2023). Our fundamental argument here is not to shift completely from one assessment paradigm (i.e., using only short discrete, static items) to another (i.e., using only extended, interactive performance tasks with resources and feedback), but rather encourage the development and use of a more diversified set of assessment experiences where the breadth and depths of tasks and associated measurement models are aligned with what the assessment intends to measure.

In this section, we present two examples of innovative assessments that connect with the five design principles discussed above: one in the context of large-scale assessment–the PISA 2025 Learning in the Digital World assessment–and the other–the Platform for Innovative Learning and Assessment–in the context of formative, classroom assessment. These two examples intend to show how the principles above can be operationalized in different contexts of use.

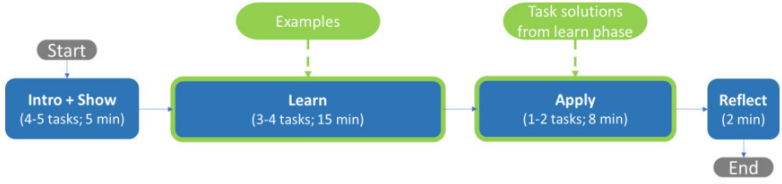
It should also be noted that the development of valid and effective innovative assessments must follow the principles of a coherent design process. Exploiting the advantages of technology requires well-designed task models that define the features of responses that matter for scoring, as well as a clear conceptual understanding of how different patterns in student behaviors align with competency states or strategies at different levels of proficiency in the target construct. The Evidence-Centred Design (ECD) framework provides a rigorous and practical framework for making coherent design decisions, particularly in the context of measuring complex competencies in interactive environments (Mislevy & Riconscente, 2006; Piacentini, 2023).

Large-scale assessment: PISA 2025 Learning in the Digital World

In its forthcoming 2025 cycle, the OECD's Programme for International Student Assessment (PISA) will include an assessment of Learning in the Digital World (LDW). The LDW assessment intends to provide comparable data on students' readiness to learn and problem solve using computational tools in digital learning environments. The assessment is based on a constructivist notion of learning, i.e., that learners are active participants in building their knowledge, and that knowledge and understanding are built through interactions between other people or objects. It is thus aligned with contemporary theories of cognition and aims to capture knowledge-in-use, and particularly, students' capacity to solve problems and self-regulate their learning. To distinguish more effective learners from less effective learners, the assessment provides opportunities for students to engage in knowledge construction activities. Depending on the unit context, students can use computational tools (simple programming, modelling or simulation environments) to build and represent their emergent understanding.

The assessment is structured as a learning experience (design principle 1). This diverges significantly from the traditional PISA format (a series of stimuli together with short, independent items) to a series of connected lessons within extended units of 30 minutes, organized as a series of phases (See Figure 2). First, a virtual agent introduces the overall learning goal of the unit and frames the experience as a tutoring session (e.g., "I'm going to teach you how to…"). Students then complete a pre-test battery of short, discrete items that aim to measure their prior knowledge of the concepts and practices they will learn and apply in the unit. This pre-test is intended to serve as useful information for contextualizing student behaviors and performance throughout the unit (design principle 2). The students then complete a tutorial to familiarise themselves with the core functionalities and affordances of the learning environment before spending approximately 10–15 minutes in the learning phase of the unit. Here, students complete a series of carefully scaffolded tasks that each focus on a concept or practice in the context of a simple to moderately open problem. In the final "challenge" phase of the unit, students must apply the concepts and practices they have learned throughout the unit to solve an open, complex and multi-step problem. This design aims to immerse the students in an authentic, digital learning experience following a sequence of coherent instructional tasks, starting from basic and progressing towards more complex ("low floor, high ceiling") applications (design principle 5). Two detailed prototype units are included in the assessment framework to illustrate this approach (OECD, 2023).

Figure 2
Unit structure in the PISA 2025 Learning in the Digital World test
Source: OECD (2023).



Part of the LDW assessment construct relates to students' capacity to engage in self-regulated learning (SRL), which refers to the monitoring and control of one's metacognitive, cognitive, behavioral, motivational and affective processes while learning (Panadero, 2017). From a design perspective, generating observations on SRL processes requires giving students problems that can only be solved through multiple iterations, providing them with information resources, and developing tools that students can use to monitor and evaluate their progress as well as receive feedback (design principles 3 and 4). Several affordances were embedded in the test environment for these purposes (see Roll & Barhak-Rabinowitz, 2023, for a detailed presentation). For example, students can access worked examples to help them solve a task; they can receive feedback by testing their computational artifacts (e.g., programs or models) against expected outcomes or by asking the tutor to check their work; and they can choose to view correct solutions after submitting a response to each task. At the end of each unit, students are also asked to evaluate their performance and report the effort they invested and the emotions they experienced as they worked. The assessment thus integrates the idea that we can better measure complex socio-cognitive constructs by giving students choice in the assessment, and monitoring not just how well students ultimately solve problems but also how they go about learning to do so.

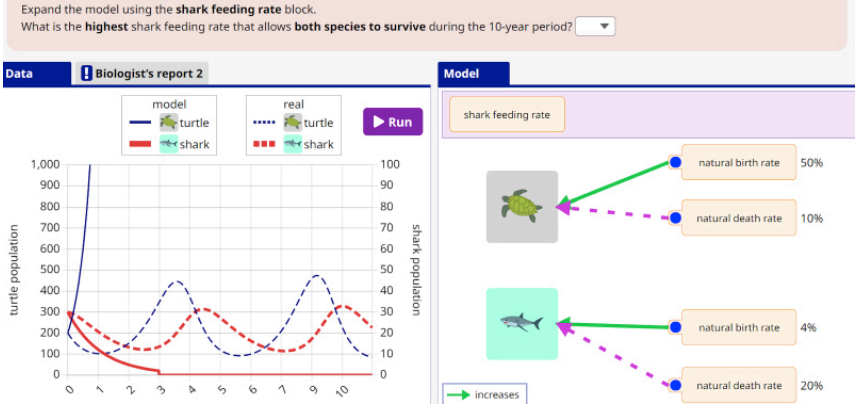
One of the main design challenges for this large-scale assessment centered on how to define learning resources and feedback that could be standardized and automated, given the heterogeneous student population and limited testing technology. This was challenging for several reasons. Effective feedback is both task- and tutee-dependent: it must be based on a complete model of the task's demands, affordances, and solution space, while also adapting to the performance

of the student. Without attending to both, the system cannot generate feedback that is useful for all students, in turn failing to bring each student to their zone of proximal development (Vygotsky, 1978). Artificial Intelligence holds promise for answering to this challenge, at least for some types of learning and assessment experiences (see Hu et al., 2023), but this was not practically, politically, or financially viable to integrate in the PISA 2025 cycle. In the LDW assessment, personalized feedback is provided implicitly through experimentation tools, i.e., generated by the system in response to students' actions (e.g., testing their model), or explicitly by the virtual agent; and more general feedback is available to students by accessing complete solutions upon task completion. For example, in the released unit Conservation[2], students construct a model of a marine ecosystem that can be used to predict what will happen in the future under different scenarios (See Figure 3). They are instructed that their model is accurate when its predictions (the straight lines) correspond to historical data (the dotted lines). Each time they run their model they receive visual feedback and are expected to continue improving their model parameters until they obtain a good fit with the real data.

Figure 3

Excerpt from the PISA 2025 LDW released unit "Conservation"

Source: OECD (2023) and accessible online at https://conservation.netlify.app.



2 Interactive example experience accessible at: https://conservation.netlify.app.

Designing effective yet standardized (i.e., non-adaptive) learning resources represented an even greater challenge. On the one hand, resources needed to be useful enough to most students so that they were incentivized to consult them and so that they provided useful hints or guidance on optimal strategies. On the other hand, resources needed to encourage the transfer of concepts without an excessive cognitive demand or without giving students prescriptive solution steps. Experts decided to address this complex trade-off by developing worked examples that students could access for each task. The worked examples provide an indication of how similar problems can be solved but students must transfer the concepts/practices described in the example to a different problem context.

The evidence needed to make claims on students' proficiency in the target knowledge, skills and attitudes of the assessment is derived from multiple sources. These include students' responses to explicit questions in the learning and challenge phase of the units, the quality and completeness of the computational artifacts (i.e., programs and models) they produce, and sequences of process data (log files) from their interactions with the environment and its affordances. The significant use of process data for scoring in the LDW assessment represents another innovation in large-scale assessment. Some uses of process data for scoring are relatively straightforward: for example, in modeling units, process data can indicate whether students completed all the steps required in constructing their model (e.g., conducting sufficient experiments to make an evidence-based conclusion on the relationship between two variables). Using process data is also essential–but much trickier–for evaluating students' self-regulated learning processes, given that any potential evidence of these behaviors must be evaluated in the context of students' previous and subsequent actions as well as the state of the learning environment when the action takes place. For example, the action of seeking help from the virtual agent may only be considered evidence of monitoring one's progress and adapting if the student actually needs help (i.e., the student is stuck) and if they implement what the tutor recommends (where this is possible). Detailed evidence rules have been developed to describe how student response and response process data should be interpreted for scoring (OECD, 2023; Foster et al., under review). These evidence rules will also be complemented by applying data mining methods to pilot and field trial data to uncover other potential evidence of coherent self-regulated learning behaviors.

Taken together, the innovations presented here for the LDW assessment align with our five design principles and respond to well-defined evidentiary needs. As further elaborated by Piacentini (2023), the assessment intends to respond to three interconnected questions: 1) what types of computational problems can students solve?; 2) to what extent are they able to learn new concepts in this domain by solving sequences of connected, scaffolded tasks?; and 3) to what extent is this learning supported by productive behaviors and learning processes, such as using resources or monitoring progress towards their learning goals? Following an ECD process, these questions have guided the domain analysis and domain modeling, defined the cognition (i.e., student) model of the assessment, oriented the design of tasks needed to elicit necessary observations about student behaviors and processes, and guided the assessment analysis plans. Following the assessment, the OECD intends to produce multi-dimensional reports of student performance including measures of their: 1) overall computational problem-solving competence (represented on a scale, as in other PISA assessments); 2) learning gains, i.e., the extent to which students' can learn and successfully apply concepts and practices during the experience; and 3) self-regulated learning behaviors. The goal is to provide policy makers with actionable information beyond a single score and position in an international ranking, including more nuanced descriptions of what students can do and cannot do, and the aspects of their performance that deserve more attention.

Formative assessment: The Platform for Innovative Learning Assessments (PILA)

The Platform for Innovative Learning Assessments (PILA) is a research laboratory coordinated by the OECD[3]. Its aim is to design learning and assessment experiences that provide real-time feedback on student performance, typically for use in the context of classroom instruction. One overall objective of PILA is to make assessment designers, programmers, measurement experts, and educators work together to explore new ways to close the gap between learning and assessment.

One of the assessment applications developed in PILA called Karel focuses on computational problem solving[4]. Students use a block-based visual programming

3 https://pilaproject.org/

4 Several tasks in the computational problem-solving application are publicly available as demos, accessible here: https://demo-gallery-karel.netlify.app/karel-player

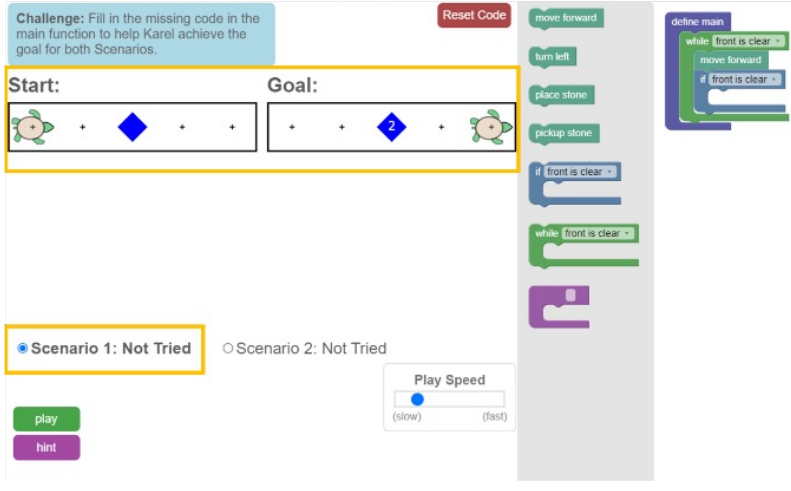
interface to instruct a turtle robot ("Karel") to perform certain actions. The tasks in the Karel application are designed to be "low floor, high ceiling" (design principle 5): low floor because the visual programming interface is intuitive enough for students who have minimal experience with programming and digital tools to still make progress, and high ceiling because the tasks are open-ended and allow for higher-performing students to challenge themselves by building "expert" solutions. programming and digital tools to still make progress, and high ceiling because the tasks are open-ended and allow for higher-performing students to challenge themselves by building "expert" solutions.

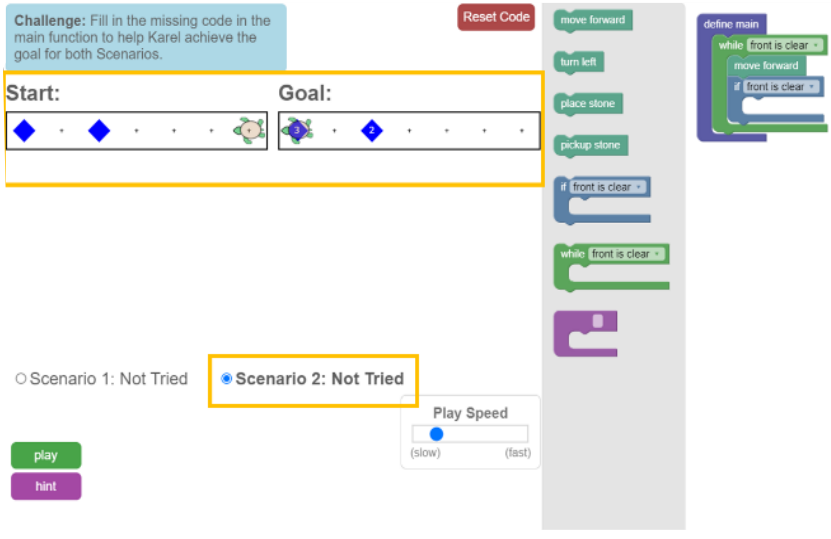
Figure 4 shows an example of a low floor, high ceiling task asking students to create a single program that moves Karel to the goal state shown in two different scenarios. To solve the problem, students can toggle between the two scenarios to visually observe the differences in the environment and how well their program works in both scenarios. Even students with solid programming skills generally require multiple iterations before finding an optimal solution for both scenarios, but the scoring models for the task take into account partial solutions (e.g., solving the problem in only one scenario).

Figure 4

Example of a low floor, high ceiling task in the Karel PILA application

Source: Platform for Innovative Learning Assessments (PILA). https://pilaproject.org/





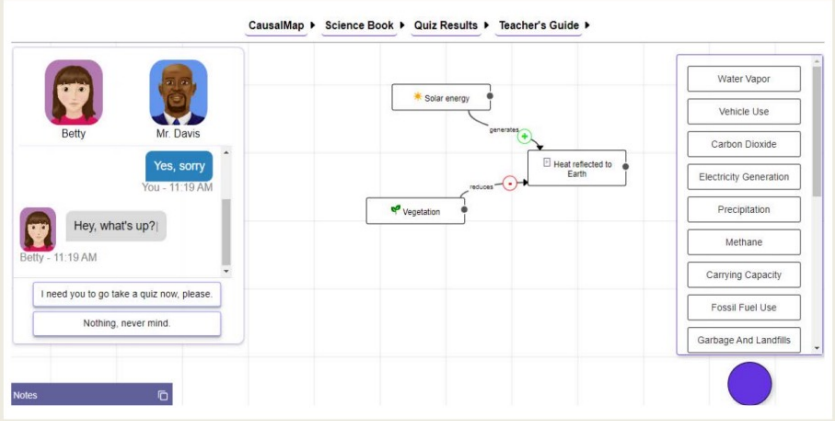
PILA includes a series of assessment experiences designed by experts and that are structured as a progression of increasingly complex tasks with a common learning target (e.g., using functions efficiently)–thus effectively creating an extended performance task (design principle 1). As they progress through the sequence of tasks, every student will experience a point in which they struggle and need to make multiple attempts to succeed. The sequence is designed so that students are expected to transfer what they have learned in earlier tasks to more complex tasks. In the near future, PILA plans to include adaptive sequence pathways (i.e., problem sequences that adapt in real time to student performance) to further align the assessment experience with students' previous knowledge and skills (design principle 2). A related and active research line in PILA consists of developing statistical models that can accumulate information on students' current levels of knowledge and skills from all the tasks they complete in the platform. For example, when a student successfully completes a programming problem in the Karel application, the estimate of her ability in programming is updated if the problem is considered relatively difficult given her known ability level.

Instructional support and opportunities for students to receive feedback on their work are integrated as key features of all PILA applications (design principles 3 and 4). In the Karel application, students receive visual feedback on their program by testing and comparing the real state with the goal state. If students cannot make progress in the task they can also access learning resources, such as hints or "good" and "bad" example solutions to similar problems. In another PILA application called Betty's Brain, developed with researchers at Vanderbilt University (see Biswas et al., 2015), students engage in an extended learning-by-teaching task in which they teach a virtual agent, "Betty", about a given scientific phenomenon5. They do so by searching through hyperlinked resources in a digital textbook and constructing a concept map that represents their emergent understanding of the phenomenon[5] in question (See Figure 5 for an example task focused on climate change). Students can ask Betty to take tests using the information represented in the concept map; Betty's performance on the test then informs students if they have any wrong or missing elements in the concept map. To be successful in the environment, students must apply metacognitive strategies for setting goals, developing plans for achieving these goals, monitoring their plans as they execute them, and evaluating their progress. They have access to learning resources (the science book) and can seek feedback on their emergent knowledge and understanding (by asking Betty to take a quiz, or asking another virtual agent, "Mr Davis", for help). The data collected in this blended learning and assessment environment are thus particularly valuable for evaluating students' self-regulated learning skills.

5 An extended Betty's Brain demo task focused on learning about ecosystems is publicly available here: https://new-kl-package.bettysbrain.pages.dev/practice/resource-library

Figure 5
Excerpt from the Betty's Brain PILA application

Source: Platform for Innovative Learning Assessments (PILA). https://pilaproject.org/



Several other PILA applications integrating our five design principles are currently under development. Some of them focus on foundational knowledge and skills (mathematics, reading and scientific inquiry), while others target transversal skills (game design, social problem solving). Digital technologies, such as AI avatars that converse with students in natural language or tools that support collaborative problem-solving between students, are also being integrated to collect more quality evidence on complex competencies, such as collaboration and communication, in different contexts.

Assessment systems like PILA hold the promise of reducing the current separation that exists between formative and summative assessments. As students complete more and more assignments in different PILA applications, they learn important concepts and skills. At the same time, PILA collects data on what they can and cannot do in different contexts, and through a principled design we can use this evidence to construct progressively more precise inferences on their competencies across multiple domains, including difficult-to-measure competencies like self-regulation. Whether these continuous systems of assessment will eliminate the need for summative assessments is an open question that is worth exploring.

Concluding remarks and reflections for future work

This chapter has argued that the next-generation of assessments should focus on observing and interpreting how students solve complex problems and learn to do new things. Reliance on traditional tests risks encouraging a skill-and-drill education system that does not prepare students adequately for future learning. Assessments often shape and influence the teaching and learning that takes place within education systems. If we want to shift policy into practice, we must be able to assess the competencies that students need for their future, meaning we must reproduce the features of these learning situations in assessments.

To do this, we need a good understanding of what drives productive learning and therefore what students should be expected to demonstrate in such situations. Research from the learning sciences concludes that experts (i.e., those that can use their knowledge to solve new problems) acquire competence through interactions in communities of practice, organize their knowledge into schema they continuously consolidate, and overcome the limits of their current knowledge through self-regulation. These processes can be reproduced and observed, at least to some extent, with extended assessment tasks that stimulate students to engage productively with learning resources and affordances in relatively open environments. Research on instructional design points towards some general design innovations that are worth trialing at scale in assessment, including the use of interactive and extended performance tasks that confront students with problems they have not seen before, that provide them with resources to explore and build upon their understanding, that assist them with feedback and support when they struggle to make progress, and that adapt in terms of complexity to what students can and cannot do.

Clearly, creating these types of assessment experiences is not without complexity from a design perspective; and making sure that they work from a measurement perspective is even more challenging. It will require coordinated efforts in multiple directions. First, we need a solid understanding of how learning unfolds across different domains and how to design tasks that reproduce the key elements of the social processes behind developing expertise in those domains. Second, we need improved measurement models that can validly interpret complex patterns of behaviors in dynamic environments, where both a student's knowledge and the problem state change as a result of their actions. Third, we need enhanced

processes to validate the inferences we make from these more complex and performance-oriented assessments, including their sensitivity to cross-cultural differences.

The two examples presented in this chapter–the OECD's PISA LDW test and PILA platform–represent initial engagement with the enterprise of innovating assessments. Yet no single assessment can evaluate all forms of knowledge and skill that we value for students; nor can a single instrument meet all of the educational goals held by parents, practitioners and policymakers. Rather, we should aim to establish a coordinated system of assessments in which different tools are used for different purposes, but where all assessments faithfully represent the constructs of interest and model good teaching and learning practices (Pellegrino, 2023). In these systems, assessments should measure higher-order cognitive skills (i.e., those that support knowledge-in-use and 21st-century competencies) in authentic and realistic contexts, should be instructionally valuable and provide instructionally useful information about the extent to which students are capable of performing such tasks, and should be valid, reliable and fair for a range of learners (Darling-Hammond et al., 2013; Pellegrino, 2023). The major challenge is determining the conditions and resources needed to create coherent systems of assessments that work across different contexts (i.e., from classrooms to larger organizational units, such as districts, states, countries, etc.), that meet the ambitious goals we have for the educational system, that meet the information needs of different stakeholders, and that align with the criteria above.

Aspects of this assessment system design and implementation challenge are further discussed in the concluding chapter of *Innovating Assessments to Measure and Support Complex Skills* (Pellegrino et al., 2023). What is clear is that the work envisioned in that report, and summarised in this chapter, cannot be undertaken without an investment of multiple forms of capital: intellectual, financial and political. Next-generation assessment development is inherently a multidisciplinary enterprise involving different communities of experts (e.g., assessment designers, technology developers, learning scientists, domain experts, measurement experts, data scientists, educational practitioners, policy makers) that need to work together collaboratively to find solutions to the many conceptual and technical challenges that come with designing innovative assessments of complex competencies.

The development of assessments for application and use at any reasonable level of scale is a time consuming and costly enterprise, with the bulk of the substantial funds currently expended at national and international levels on assessment programs focused on traditional disciplinary domains that fall within conventional parameters for task development, delivery, data capture, scoring and reporting. Substantial fiscal capital is required to assemble and support the multidisciplinary teams needed to conduct research and development supporting the creation of innovative next-generation assessments, as well as for efforts to bring them to full maturity by scaling up their implementation when evidence exists that they can effectively address the challenge of measuring the constructs that matter. Finally, it is hard to make major changes within existing assessment systems when there are well-established operational programs that are entrenched in practice and policy. Change of the type needed requires strong political will and vision to encourage people to think beyond what is possible now or even in the near future– and that includes policy makers, but also the educational assessment development community, the measurement and psychometric community, and the educational practice community. Without political will, it will be impossible to generate sufficient fiscal capital to assemble the intellectual capital required to pursue next-generation assessment development and implementation, and in turn achieve meaningful change in educational assessment.

To illustrate these points, the development of the LDW assessment was only possible because of the convergence of these three types of capital. The innovative assessment included in each PISA cycle is now seen as a safe space to test important innovations in task design and analytical models that can be transferred to other assessments or that can provide inspiration for the development of national assessments once their value is proven. The PISA Governing Board provided the financial and political support needed to start the development of the test several years before the main data collection and to engage in extensive validation processes for its design and analytical choices. Further resources were also made available by research foundations that recognized the value of innovating assessments. The development of the assessment has also been steered by a group of experts with different disciplinary backgrounds: subject matter experts worked side-by-side with psychometricians, scholars in learning analytics and experts in UI/UX design. This cross-fertilisation was important to make space for new methods of evidence identification in digital learning environments while

keeping in mind the core objective to achieve comparable metrics that result in valid interpretations of performance differences across countries and student groups.

In summary, we argue that we need many new disciplinary and cross-disciplinary assessments to provide an exhaustive description of the quality of educational experiences and the extent to which students are prepared to thrive in their futures. There is evidence that innovative assessments of educationally and socially significant competencies are both desirable and possible, and we have described some overarching characteristics of what these assessments should look like. The evidence also suggests that cooperation and collaboration on a global scale, including through the design of open-source technology and model tasks, may well be the best and only way to achieve such advances—at least for the time being.

References

Azevedo, R., & Aleven, V. (Eds.). (2013). International handbook of metacognition and learning technologies. *Springer.* https://doi.org/10.1007/978-1-4419-5546-3

Binkley, M., Erstad, O., Herman, J., et al. (2011). Defining twenty-first century skills. In Griffin, P., McGaw, B., & Care, E. (Eds.), *Assessment and teaching of 21st-century skills.* Springer. https://doi.org/10.1007/978-94-007-2324-5_2

Biswas, G., Segedy, J., & Bunchongchit, K. (2015). From design to implementation to practice a learning by teaching system: Betty's Brain. *International Journal of Artificial Intelligence in Education, 26*(1), 350–364. https://doi.org/10.1007/s40593-015-0057-9

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2019). Implications for educational practice of the science of learning and development. *Applied Developmental Science, 24*(2), 97–140. https://doi.org/10.1080/10888691.2018.1537791

Darling-Hammond, L., Herman, J., & Pellegrino, J., et al. (2013). *Criteria for high-quality assessment.* Stanford Center for Opportunity Policy in Education.

DiCerbo, K., Shute, V., & Kim, Y. (2017). The future of assessment in technology-rich environments: Psychometric considerations. In Spector, J., Lockee, B., & Childress, M. (Eds.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy.* Springer International Publishing. https://doi.org/10.1007/978-3-319-17727-4_66-1

Dumont, H., Istance, D., & Benavides, F. (Eds.). (2010). *The nature of learning: Using research to inspire practice.* OECD Publishing. https://doi.org/10.1787/9789264086487-en

Ericsson, K. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In Ericsson, K., Charness, N., Feltovich, P., & Hoffman, R. (Eds.), *The Cambridge handbook of expertise and expert performance.* Cambridge University Press. https://doi.org/10.1017/cbo9780511816796.038

European Commission (2019). *Key competences for lifelong learning*. Publications Office of the European Union. https://data.europa.eu/doi/10.2766/569540

Fadel, C., & J. Groff (2018). Four-dimensional education for sustainable societies. In Cook, J. (Ed.). *Sustainability, human well-being, and the future of education.* Springer International Publishing. https://doi.org/10.1007/978-3-319-78580-6_8

Foster, N., Holmes, J., Linsenmayer, E., Zoanetti, N., & Yildiz, H. (under review). Developing rule-based scoring methods that capture student progress in computational problem solving in open, interactive tasks. *Learning and Individual Differences*.

Foster, N. (2023). 21st-century competencies: Challenges in education and assessment. In Foster, N., & Piacentini, M. (Eds.), I*nnovating assessments to measure and support complex skills*. OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Foster, N., & Piacentini, M. (Eds.). (2023). *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Guzdial, M., Rick, J., & Kehoe, C. (2001). Beyond adoption to invention: Teacher-created collaborative activities in higher education. *Journal of the Learning Sciences, 10*(3), 265–279. https://doi.org/10.1207/s15327809jls1003_2

Hatano, G. (1990). The nature of everyday science: A brief introduction. *British Journal of Developmental Psychology, 8*(3), 245–250. https://doi.org/10.1111/j.2044-835x.1990.tb00839.x

Hatano, G., & Oura, Y. (2003). Commentary: Reconceptualizing school learning using insight from expertise research. *Educational Researcher, 32*(8), 26–29. https://doi.org/10.3102/0013189x032008026

Hu, X., Shubeck, K., & Sabatini, J. (2023). Artificial Intelligence-enabled adaptive assessments with Intelligent Tutors. In Foster, N., & Piacentini, M. (Eds.), *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences, 21*(1), 45–83. https://doi.org/10.1080/10508406.2011.591717

Kyllonen, P. C., & Sevak, A. (2024). *Charting the future of assessments.* ETS Research Institute. https://www.ets.org/Rebrand/pdf/FoA_Full_Report.pdf

Loibl, K., Roll, I., & Rummel, N. (2016). Towards a theory of when and how problem solving followed by instruction supports learning. E*ducational Psychology Review, 29*(4), 693–715. https://doi.org/10.1007/s10648-016-9379-x

Marion, S. F., & Pellegrino, J. W. (2007). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice, 25*(4), 47–57. https://doi.org/10.1111/j.1745-3992.2006.00078.x

Mislevy, R. (2018). *Sociocognitive foundations of educational measurement.* Routledge.

Mislevy, R., & Riconscente, M. (2006). Evidence-centered assessment design. In Downing, S., & Haladyna, T. (Eds.), *Handbook of test development.* Lawrence Erlbaum.

Nunes, C., Nunes, M., & Davis, C. (2003). Assessing the inaccessible: Metacognition and attitudes. *Assessment in Education: Principles, Policy & Practice, 10*(3), 375–388. https://doi.org/10.1080/0969594032000148109

OECD (2023), *PISA 2025 Learning in the Digital World assessment framework (second draft).* https://www.oecd.org/content/dam/oecd/en/topics/policy-sub-issues/learning-in-the-digital-world/

OECD (2018). *The future of education and skills 2030.* OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2018/06/the-future-of-education-and-skills_5424dd26/54ac7020-en.pdf

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology,* 8, https://doi.org/10.3389/fpsyg.2017.00422

Pellegrino, J. (2023). Introduction: Arguments in favour of innovating assessments. In Foster, N., & Piacentini, M. (Eds.). *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Pellegrino, J., Chudkowsky, N., & Glaser, R. (Eds.). (2001). K*nowing what students know: The science and design of educational assessment.* National Academy Press. http://faculty.wiu.edu/JR-Olsen/wiu/common-core/precursor-documents/KnowingWhatStudentsKnow.pdf

Pellegrino, J., DiBello, L., & Goldman, S. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550

Pellegrino, J., Foster, N., & Piacentini, M. (2023). Conclusions and implications. In Foster, N., & Piacentini, M. (Eds.). *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Pellegrino, J., & Hilton, M. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st-century.* National Academies Press. https://doi.org/10.17226/13398

Piacentini, M. (2023). Defining the conceptual assessment framework for complex competencies. In Foster, N., & Piacentini, M. (Eds.), *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Piacentini, M., Foster, N., & Nunes, C. (2023). Next-generation assessments of 21st-century competencies: Insights from the learning sciences. In Foster, N., & Piacentini, M. (Eds.), *Innovating assessments to measure and support complex skills. OECD Publishing.* https://doi.org/10.1787/e5f3e341-en

Roll, I., & Barhak-Rabinowitz, M. (2023). Measuring self-regulated learning using feedback and resources. In Foster, N., & Piacentini, M. (Eds.), *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Roll, I., Aleven, V., McClaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction, 21*(2), 267–280. https://doi.org/10.1016/j.learninstruc.2010.07.004

Roll, I., Holmes, N. G., Day, J., & Bonn, D. (2012). Evaluating metacognitive scaffolding in guided invention activities. I*nstructional Science, 40*(4), 691–710. https://doi.org/10.1007/s11251-012-9208-7

Sabatini, J., Hu, X., Piacentini, M., & Foster, N. (2023). Designing innovative tasks and test environments. In Foster, N., & Piacentini, M. (Eds.), *Innovating assessments to measure and support complex skills. OECD Publishing.* https://doi.org/10.1787/e5f3e341-en

Schwartz, D., & Arena, D. (2013). *Measuring what matters most: Choice-based assessments for the digital age.* The MIT Press. https://doi.org/10.7551/mitpress/9430.001.0001

Schwartz, D., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics education. *Cognition and Instruction, 22*(2), 129–184. https://doi.org/10.1207/s1532690xci2202_1

Simon, H. (1980). Problem solving and education. In Tuma, D., & Reif, R. (Eds.), *Problem solving and education: Issues in teaching and research.* Erlbaum.

Scalise, K., Malcolm, C., & Kaylor, E. (2023). Analysing and integrating new sources of data reliably in innovative assessments. In Foster, N., & Piacentini, M. (Eds.), *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Scott, C. (2015). The futures of learning 2: What kind of learning for the 21st-century? Education, Research and Foresight: Working Papers. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000242996

Shute, V., Hansen, E., & Almond. R. (2008). You can't fatten a hog by weighing it—or can you? Evaluating an assessment for learning system called ACED. I*nternational Journal of Artificial Intelligence in Education, 18*(4), 289–316.

Shute, V., & Kim, Y. (2013). Formative and stealth assessment. In Spector, J., et al. (Eds.), *Handbook of Research on Educational Communications and Technology (4th Edition).* Lawrence Erlbaum.

Taylor, J. L., Smith, K. M., van Stolk, A. P., & Spiegelman, G. B. (2010). Using invention to change how students tackle problems. *CBE—Life Sciences Education, 9*(4), 504–512. https://doi.org/10.1187/cbe.10-02-0012

Timmis, S., Broadfoot, P., Sutherland, R., & Oldfield, A. (2016). Rethinking assessment in a digital age: Opportunities, challenges and risks. *British Educational Research Journal, 42*(3), 454–476. https://doi.org/10.1002/berj.3215

Thornton, S. (2012). Issues and controversies associated with the use of new technologies. In Gormley-Heenan, C., & Lightfoot, S. (Eds.), T*eaching Politics and International Relations*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137003560_8

van Joolingen, W., de Jong, T., Lazonder, A. W., Savelsbergh, E. R., & Manlove, S. (2005). Co-Lab: Research and development of an online learning environment for collaborative scientific discovery learning. *Computers in Human Behavior, 21*(4), 671–688. https://doi.org/10.1016/j.chb.2004.10.039

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Wiske, M. (Ed.). (1997). *Teaching for understanding: Linking research with practice.* Jossey-Bass.

World Economic Forum (2015). New vision for education: Unlocking the potential of technology. https://www3.weforum.org/docs/WEFUSA_NewVisionforEducation_Report2015.pdf

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430–1459. https://doi.org/10.1002/tea.21658

Designing for the Future: Toward an R&D Agenda to Promote Inclusive, Human-Centered Assessment Systems

Temple S. Lovelace, Orrin T. Murray, and Laura S. Hamilton

Abstract

This chapter examines how inclusive human-centered design (IHCD) can transform educational assessment systems to better serve diverse learners. As the demands of education have expanded over time, assessment paradigms must likewise evolve beyond narrow academic measures to support holistic student development. The authors argue that assessment systems should be intentionally designed through IHCD principles that center the experiences and needs of learners, educators, and communities. The chapter explores three critical layers of an inclusive assessment R&D infrastructure: content that reflects modern understanding of development across domains, technological capabilities that enhance assessment utility, and broader policy support that ensures coherence and equity. Drawing on the Gordon Commission's Principles for Assessment in Service of Learning, the authors provide guidance for creating assessment systems that are transparent, authentic, culturally responsive, and actionable. The chapter addresses the potential benefits and risks of emerging technologies, particularly AI, while emphasizing the importance of stakeholder engagement in assessment design. By reimagining assessment through IHCD, education systems can move beyond current incoherent approaches toward paradigms that genuinely support learning, equity, and the multifaceted purposes of public education in preparing young people for education, careers, and civic life.

Introduction

The knowledge that an adolescent might have needed to successfully navigate young adulthood in the United States 150 years ago pales in comparison to what an adolescent must know and have faculty with today. And while the broad outlines might look familiar, what a school must do to prepare its students to succeed has significantly expanded in the last century and a half (Roschelle, Pea, Hoadley, Gordin & Means, 2000). Although the early struggles over the purpose of public schools have shifted away from the battlegrounds surrounding religion and morals, other social, cultural, and political concerns, including topics of race, gender, and social and emotional development, continue to roil the system of public education in the United States (Lampen, 2022; Mehta, 2025; Schwartz, 2021).

Perhaps the most consequential contemporary shift in how people talked about the purpose of education began with the publication of *A Nation at Risk,* which signaled a new front in the battle over the purpose of public schools (National Commission on Excellence in Education, 1983). As the United States' economic footing shifted from a manual labor focused workforce—rooted in manufacturing and agriculture—toward one that relied more on cognitive competencies (with an emphasis on technical skills in math and science) the last 50 years have seen increased emphasis on skills that directly supports this newer economic world standing (Hanushek & Woessmann, 2012; Zaber et al., 2019). For several decades, the idea that public schools serve the explicit purpose of developing vibrant citizens—people who vote in every eligible election, serve on juries, run for public office and are active at all levels in their communities—was at best implied, and, at worst largely forgotten by those charged with managing our education systems. Recent years have witnessed a renewed emphasis on cultivating citizens and promoting holistic development (Lee, White, & Dong, 2023; Vinnakota, 2019), but the purpose of public education remains contested, and current assessment and accountability policy does not, for the most part, reflect this broader vision (Hamilton & Martinez, 2024).

In keeping with the views and evidence presented throughout this chapter and volume, we assert that publicly funded schools must prepare young people to succeed and thrive in education settings (i.e., PreK–16+), careers, and civic and community life—though we acknowledge that schools cannot accomplish this on their own. These are ambitious objectives for education, and how we monitor and seek to address opportunities to reach these objectives—with the explicit support of assessment paradigms—must be grounded in the multifaceted purposes of

public education. Moreover, we argue that these assessment paradigms—many which do not currently exist—must rest on inclusive and human-centered designs. In offering assessment paradigms, as a bridging agent between our objectives and practices, it is critical to highlight that assessments (as with all tools or technologies) have both been designed to harm and have been used to harm in ways that were intentional and unintentional. The approach to assessments we offer here intentionally focuses this powerful tool explicitly in the service of ensuring that the benefits of assessment outweigh potential harms for all learners.

HCD and Inclusion

In some respects, the idea that the design of assessment paradigms has not been considered from the perspective of learners and educators making use of it seems nonsensical. Put another way, the concept of human-centered design (HCD), from the perspective of the words offered, appears redundant and obvious, for we have often overlooked the true meaning of HCD even as we have aimed to apply its principles. The status quo within design, prior to HCD-based movements, was elitist, and expert driven—where expertise centered on a canon cultivated by a small coterie of gatekeepers. Even as HCD gains popularity, it would be shortsighted to suggest the struggle in the design world is over; it is not. The rapid and wide adoption of HCD certainly opened the door for designers to consider the role that a more empathetic approach to inclusion, centering people first (e.g., the inclusion of Empathize in Stanford d.school's design thinking process or the Gather Information step in IDEO's process), and it has shifted the field in ways that the prior generations of designers might not recognize. For example, as the bulk of design work shifted from material objects to intangible objects like websites and applications for digital platforms, from a design perspective, a focus on beauty has shifted somewhat to focus on utility. As we consider the role of assessment paradigms, this lesson from inclusive, human-centered design can be instructive; for we must contend with what "utility" means in educational assessment and how that can impact the design, implementation, and use of the artifacts from those paradigms. For, in an inclusive and human-centered approach to assessment R&D, considering who is involved at every step is important to fully addressing the role of utility within assessment paradigms. The emphasis on utility signals that an inclusive, human-centered approach forces designers to consider the "for whom" their designs are intended for—who can, who will, and in what ways do the people who imbue them enter into the design conversation. Therefore, IHCD became an

opportunity for us to consider the question of whom does assessment design serve (e.g., "Why are we doing this? Why is it being done this way?").

As we look to HCD for developing assessments for education, interest in education addressing social justice, social and emotional development, equity and inclusion, though in retreat at the federal level in early 2025, still seems ripe for considering the questions, "for *whom* is this assessment designed?' as well as "for *what* is this assessment designed?"

The Evolution of HCD Ideologies

In *The Design of Everyday Things,* Norman describes HCD as "an approach that puts human needs, capabilities, and behavior first, then designs to accommodate those needs, capabilities, and ways of behaving." (Norman, 2013). Norman goes on to note that HCD starts with "a good understanding of people and the needs [of these people] that the design is intended to meet." What people need, implies function and use in ways that could outweigh a pure aesthetic, or even functional beauty. On its face, however, much like the Declaration of Independence or the U.S. Constitution, details about who the people whose needs, capabilities and behaviors are taken into account, matter.

The rise of what we now label human-centered design can be traced to a set of practices that helped to fuel the successes of David Kelley and his design firm, IDEO. While browsers like Netscape are seen as the principal point of leverage for popularizing the internet (McCullough 2018), the rapid expansion of user interface and user experience skills, which draw heavily on HCD, provide infinite use cases for why HCD matters to the wide use and thereby success of the internet. It is however, only recently that education researchers seriously sought to consider the concept of design.

The persistent failure of public schools to close opportunity gaps is but one telling marker for why we need to clarify that the kind of HCD for designing assessments we have in mind is explicitly inclusive. Where commercially focused designers are often rewarded to segment and focus on particular groups to the exclusion of others, such a design principle is incompatible (even if, from a practical perspective, such principles are ignored or pushed to the periphery) with compulsory public education—it is harder to pointedly exclude some students in how we think about teaching and learning, and thus assessments.

Building upon the solid foundation provided by researchers promoting the science of learning (Goldman & Lee, 2024; Osher, Cantor, Berg, Steyer & Rose, 2020), we believe that the assessment systems that are enacted in the future will need to be more responsive, relative to our current systems, to the assets and experiences students bring into learning environments. As such, these future assessment systems could be designed to offer educators timely and relevant data to inform a holistic, responsive approach to instruction. Even more, these future assessment systems should consider how caregivers and learners might also benefit from access to these data, especially as what we value as learning further extends to contexts outside of a brick and mortar school building. There are also lessons on the transformation needed from other fields. For example, the functional approaches, such as those found in medicine are another important consideration when thinking about IHCD (Cantor, 2021). A surgeon using rigid systems to assess patient risks, built on population averages will, on average, make care decisions that lead to positive patient outcomes, but treatment decisions on either side of the average will more often lead to fatal or catastrophic outcomes. And while a single instructional decision does not typically have the same effect as that of a surgeon, the accumulation of such decisions using assessment systems that are insensitive to some students mean that those students will more often than not receive inadequate instruction and be denied opportunities for professionally and personally rewarding pursuits.

In this chapter, we build on the work described in the earlier chapters to consider how to apply an inclusive human-centered approach to create and implement *systems* of assessment that center the needs of learners. We start by reviewing and elaborating upon the description of inclusive, human-centered design (IHCD) presented earlier in this volume. We then describe the key features of assessment systems, discuss how the IHCD principles can be enacted to develop coherent, learner-centered assessment systems, and briefly discuss the implications of advances in technology. We conclude by sharing some examples of promising developments in R&D that we believe will contribute to a future characterized by more inclusive, learner-focused assessment systems.

Applying ICHD to Systems of Assessment

Inclusive human-centered design (IHCD) is an approach to the design and development process that ensures that the process of design starts with an empathetic understanding of students' lived experiences and that the outcomes of design support a set of solutions that reflect their lived experiences. In a similar vein, assessment, as described by Mislevy (2019), is "situated in social contexts, shaped by purposes, and centered on students' developing capabilities for valued activities (p. 164)." The questions asked of inclusive approaches to human-centered design must also be asked of the development approaches used for assessment. If assessments are to be responsive, relevant, and provide the opportunity for reflection, then the methods and processes we use must follow the tenets laid forth for IHCD. These tenets, loosely summarized here and applied to assessment development, are that IHCD is about the inclusion of learners and educators into the process, "with whom," that it not only centers the purpose of the assessment process on the people most impacted, "for whom;" but that the context or environment in which the assessment process will be utilized matters, "for what". This process of centering these questions at the nexus of the assessment process is complex, yet it is a critical element to ensure we understand how assessment as a whole needs to evolve to better meet the needs of learners now, but certainly the learners in the future.

Before discussing the applications of specific IHCD principles to assessment and assessment systems, it is important to clarify how we are using these terms in the context of educational settings. "Assessment" can be defined as a "process of reasoning from evidence" (Foster & Piacentini, 2023, p.16) or, in a slightly more detailed description, "...the process and outcome of collecting and analyzing data to inform an interpretation, judgment, or decision about an attribute or property of an object, event, or phenomenon" (Ackerman et al., 2024, p.2). This definition applies to a variety of tools and approaches including large-scale, standardized achievement tests, curriculum-embedded assessments, and the informal data teachers collect almost constantly through their interactions with students. Additionally, by describing assessment as both a process and an outcome, the definition highlights the importance of design, development, administration, scoring, reporting, and the decisions that users make in response to results.

Regardless of the approaches used at each stage, assessment can serve various purposes including, but not limited to:

- Informing classroom instruction by providing information to educators and learners about learners' progress, assets and needs.
- Guiding learners' decisions about academic or career pathways.
- Selecting students into programs or institutions.
- Monitoring learning progress on a large-scale.
- Serving as a source of data for research that advances the empirical evidence base.
- Signaling to educators, caregivers, or other groups what kinds of learning outcomes they are expected to value and prioritize.

Of course, no single instrument can serve all these purposes effectively, and it is unwise to use a measure for a purpose other than the one (or ones) for which sufficient validity evidence has been gathered (AERA, APA, & NCME, 2014). Consequently, most young people will encounter numerous assessments, designed for a variety of purposes, each year as they make their way through the K–12 system and move into college, careers, or other pursuits.

Although some K–12 assessments are designed and validated for specific, relatively narrow purposes that do not typically influence students' broader educational experiences (for example, clinical tests to diagnose specific educational or mental health needs), many others influence what happens in classrooms and schools, whether or not they are designed for that purpose. These include assessments teachers create, curate, or modify to inform their instructional decisions; interim or benchmark assessments that districts might purchase to generate data regarding students' progress toward meeting state standards, and statewide accountability tests that are externally mandated but that often exert powerful effects on instruction (Faxon-Mills et al., 2013). Different assessments might send conflicting signals about what outcomes educators are expected to promote, and they might generate confusing or misleading data on student learning. As a result, their use can hinder educators' opportunities to provide a coherent instructional experience for students.

The concept of a *system* of assessment provides a way of thinking about how a set of assessments, designed for different purposes, might be integrated into a broad suite of tools and practices that sends aligned rather than conflicting signals to educators, students, and caregivers, and that is designed explicitly to support teaching and learning. Although there is no consensus definition of an "assessment system," nor a single clear framework to guide its development, a body of work conducted over the past two decades has converged around a vision put forth in the seminal 2001 volume, *Knowing What Students Know* (National Research Council, 2001). For this chapter, we draw on a definition offered in a recently published National Academy of Education volume called Reimaging Balanced Assessment Systems (Marion, Pellegrino, & Berman, 2024), which draws on *Knowing What Students Know* and subsequent scholarly work. The volume's editors and authors conceptualize balanced assessment systems as "intentionally designed to provide feedback to students and information for teachers to support ambitious and equitable instructional and learning opportunities" (Marion et al., 2024, p.2), with supports for high-quality formative assessment practices in classrooms along with large-scale district or state assessments that are designed to produce aggregate data for the purposes of monitoring, informing decisions about resource allocation or policy, and signaling expectations regarding instruction that will advance quality and equity. Coherence is a critical feature of this vision of balanced assessment systems, along with learner-centeredness, which we define as a relentless prioritization of improving student learning throughout all parts of the system (National Research Council, 2001; Marion et al., 2024).

Broadening our understanding of IHCD–the role of a coherent Assessment R&D Infrastructure

Efforts to make assessment paradigms more inclusive and human-centered can not start once the assessment arrives in front of an educator or learner. Given the role of IHCD in ensuring that value, effectiveness, and use are core components of assessment development, expanding its application to the full research and development (R&D) process is a critical step. Traditionally, assessment developers have engaged in R&D processes that reflected a relatively narrow range of expertise and that did not typically emphasize the role that assessments play in broader systems. In order to realize an assessment system that is inclusive and human-centered, it must be constructed on a multi-faceted infrastructure. The

infrastructure for assessment R&D is one that must consider not only the creation of innovative tools, but also the sustained use of those tools and the local and federal policy needed to shape and support it. This inclusive and human-centered infrastructure has to honor the experiences of the learners, educators, caregivers, school leaders, and others, but it must also prioritize an orientation towards the future of the learning and measurement sciences.

We offer that there are three layers within an inclusive, human-centered R&D infrastructure that are worthy of note. The first, focuses on the *substance or content* of our assessment systems, such as how they reflect a modernized view of child development across multiple domains, the degree to which they support culturally relevant experiences that promote the individualization of education, and their responsiveness to the growing evidence base stemming from advances in the learning sciences. The second layer contains the *technological capabilities* that expand the utility of those innovations, such as leveraging digital innovations (from capturing still and moving images in learning environments to using AI to support recommendations for pedagogical development or to tailor student development) and the use of multi-modal components within assessment. The last layer reflects the *broader landscape* of an IHCD-aligned assessment R&D infrastructure, which includes the local context and state and federal policy support needed to ensure a robust R&D experience, and how we creatively resource this expanded infrastructure. The position we are staking out moves past the incoherence of the current system of assessments by insisting that, going forward, assessments are anchored in an inclusive and human-centered R&D paradigm. Therefore, it will be critical that developers of assessments frame their ICHD with learner needs driving the alignment of educators, administrators and policy makers towards a more coherent and learner-focused structure.

Exploring the Role of an IHCD-Informed Infrastructure on the Learning and Measurement Sciences

The current PreK-16+ education system has long been impacted by the policies set forth for assessment and for standards-aligned learning. For decades, scholars have called for investigating a more comprehensive approach to learning that considers not only what academic content a learner should know, but also the skills that attend to their social and emotional competencies (Lovelace, McMurtry, and Kendall-Brooks, 2024). Even more, there is the consideration for how learners

may apply their skills for civic engagement, such as considering their role as change agents, and expanding the application of their skills to advocacy-related fields (Jagers et al., 2017). These considerations, which align to a call for new approaches in the learning sciences that consider the role of identity, agency, and wellbeing provide a compelling argument for an inclusive, human-centered assessment system that can support the creation and sustaining of learning environments that provide the opportunity for growth in these areas. However, to do so convincingly, the underlying assessment R&D infrastructure must be reflexive. For example, our current assessment R&D infrastructure does not equitably support the assessment of essential skills (i.e., 21st-century skills) at the same rate as skills associated with standards-based reading, math, and science -based content. Even more, there is less cohesiveness in the area of social and emotional competency assessment, with over 200 frameworks that are currently focused in this space and an insufficient evidence base to inform appropriate use of such assessments (Assessment Work Group, 2019; Berg et al., 2017; Harvard University, undated; Casillas, Roberts, & Jones, 2022). A related challenge is that most currently available assessment tools are designed to measure student learning in isolated ways, such as through separate math, reading, and social and emotional competency measures. As we consider how an IHCD infrastructure could better reflect what we know about how people learn, it will be critical to engage in R&D to advance assessment content and methods in ways that are more comprehensive and that reflect a more holistic understanding of student experience.

As we consider how we might prioritize these efforts going forward, the field must consider the relationship that assessment paradigms have to our curriculum and instruction infrastructure. The Gordon Commission's Principles suggest that this relationship is bi-directional and reflexive; one informs the other in a cyclical and continuous manner (Gordon & Pellegrino, 2013). Thus, the infrastructure to support assessment R&D must grow to support that. For example, our processes for measurement, our assessment practices and culture, as well as content, must be continually reviewed and updated to support new learning paradigms.

How Assessment Systems can Benefit from Application of the Principles for Assessment in Service of Learning

The Principles for Assessment in Service of Learning, a product of the Gordon Commission, offer a promising framework for informing both the design and implementation of assessment systems (Gordon & Pellegrino, 2013). In this section, we offer guidance drawn from these principles that we believe will help maximize the likelihood that systems of assessments are coherent and learner-centered and that they reflect the values and priorities of the groups who will be most affected by them. This guidance is not exhaustive but instead intended to highlight a small number of ideas that can inform current assessment system design efforts and future R&D, which we discuss later in this chapter.

Engage educators, learners, and caregivers in assessment system design. Principle 1 emphasizes the importance of transparency, which is relevant not just to assessments themselves but to the broader systems in which assessments are embedded. An inclusive approach to informing assessment system design should provide members of affected groups—mainly educators and learners, but also administrators, caregivers, and other community members—opportunities to inform all aspects of the system. This is not to suggest that professional assessment expertise is not needed or that members of these groups would make all decisions. Rather, application of Principle 1 would involve engaging with groups in appropriate ways, such as by inviting and incorporating input on initial design decisions and providing multiple opportunities for review and feedback on interim products. Engagement opportunities should be customized to the needs and assets of each group and should incorporate specific strategies and mechanisms for reviewing and addressing feedback (Hamilton & Landl, 2023). Frameworks associated with design-based research and user-focused R&D, which emphasize iterative cycles of testing, feedback, and adaptation, can be especially valuable (Armstrong, Dopp, & Welch, 2020; Design Based Research Collective, 2003; Hamilton & Murray, 2023; Schneider, 2023). Importantly, such engagement must go beyond giving educators or other groups "voice" and instead must allow these groups to exert agency and influence in ways that are meaningful and responsible.

Incorporate authentic activities into assessment design. One of the reasons efforts to adopt balanced assessment systems have often fallen short of their goals is that they have been overly influenced by externally mandated, high-stakes testing. The Elementary and Secondary Education Act (ESEA) has relied on statewide assessments as a key policy lever, resulting in tests that have led to a narrowing of curriculum and instruction, an overly constrained notion of what constitutes "effective" instruction, and a tendency for educators to rely heavily on commercially available interim or benchmark testing systems that aren't always designed to support high-quality teaching (Evans, 2024). Systems that address principle 2, which refers to progress, outcomes and processes that are transferable, must prioritize assessment tools and practices that do not always lend themselves to large-scale administration. These might include more sophisticated items and tasks such as scenario-based or portfolio assessments, along with informal assessment approaches that skilled teachers use to gauge their students' understanding and needs through their interactions in learning environments.

This is not to suggest that large-scale assessments cannot play a valuable role in systems of assessment. The National Assessment of Educational Progress (NAEP), for instance, provides a mechanism for monitoring performance and progress on a large scale and for exploring relationships between student achievement and other factors such as economic or instructional resources. Large-scale assessment data can be especially powerful when connected to information about students' learning opportunities and access to resources (National Academies of Sciences, Engineering, and Medicine, 2019). A set of assessments for monitoring progress at the state or district level could play an important role in an assessment system but must be designed and implemented in ways that prevent such tests from having significant influence over instruction, such as by refraining from calculating individual student scores or imposing consequences based on performance. In this way, large-scale tests can serve complementary purposes to other assessments in the system.

Anchor systems in whole-child conceptualizations of learning. A large and growing body of evidence on how people learn has advanced the assessment field's understanding of the inseparability of academic, social, and emotional development and the ways in which learning is influenced by one's cultural context and assets (Goldman & Lee, 2024). Educators have largely endorsed the goals of whole-child education (Hamilton et al., 2024; Rikoon et al., 2024). This

science-based perspective can help assessment systems align to principle 3, which acknowledges the importance of enabling learners to draw on their own knowledge and experiences while participating in assessment, and principle 4, which calls for assessment that supports multiple aspects of learners' development. Advances in classroom assessment tools and practices offer opportunities to incorporate a whole-child perspective. For instance, some digital assessment tools provide data not just on academic learning but on engagement and other related constructs (Linzarini & da Silva, 2024). More holistic approaches to measuring student progress in large-scale assessment systems are also being explored by states, districts, and advocacy groups (Domaleski et al., 2023; Hamilton & Martinez, 2024; National Urban League & UNIDOS, 2022). Of course, adopting a whole-child perspective carries some risks and must be undertaken thoughtfully. For instance, incorporation of social and emotional competency measurement could lead to inappropriate uses of such data and to a reliance on measures and construct definitions that are not culturally responsive to all groups (Assessment Work Group, 2019; Jagers et al., 2018). An IHCD approach provides one bulwark against misguided efforts by helping to ensure broad input on what aspects of learners' development should be measured, how they should be measured, and how the resulting data should be used.

Recognize how assessment systems can communicate and model expectations. The theory of standards-based reform, which informed the design of state accountability systems over the past few decades, emphasized the idea that clearly specified content standards could help educators understand what student learning outcomes they should promote but without offering prescriptions regarding pedagogy (Smith & O'Day, 1991). Research has shed light on ways in which this theory of action has failed to be realized and has found that it is often the tests rather than the standards that drive instruction (Koretz, 2008; Koretz & Hamilton, 2006). The fundamental problem with this view is that large-scale assessments are designed to accommodate broad content coverage in the context of limited administration time and funding, which in turn prioritizes item types that are poor models of instruction (Darling-Hammond & Adamson, 2010). An assessment system aligned with principle 5, which calls for assessments that can serve as models of learning, would include classroom assessment tools and practices that engage students in activities that reflect important, real-world activities.

Monitor and improve system quality based on the specific purposes of each component of the system. Principle 6 emphasizes the importance of clearly specifying the purpose of an assessment and ensuring that it supports appropriate inferences relevant to that purpose. Additionally, this principle echoes widely accepted assessment guidance regarding the need for multiple sources of evidence (AERA, APA, NCME, 2014). Systems of assessment typically incorporate several components, each of which is designed to serve a specific purpose (e.g., informing instruction, monitoring equity on a large-scale). Those who design and implement assessment system components need to engage in a process of continuous review and improvement to ensure that decisions based on each component are appropriate and warranted based on the quality of evidence those components produce and that the system as a whole generates credible, coherent data to inform improvement. Application of IHCD to this continuous improvement process involves seeking feedback from all relevant groups to understand the inferences and decisions they are making based on results and to elicit their recommendations for modifications.

Such an approach also helps advance principle 7, which emphasizes the need for assessment results to be tied to actionable feedback that informs decisions. Systems of assessment are typically intended to inform a variety of decisions, and it is critical that users of information from such systems understand how to make sense of information from different components of the system and how to use this information to make decisions that improve teaching and learning. This means being clear on what kinds of decisions are typically not warranted, such as the use of large-scale, standardized test scores to inform instruction (Marion, Pellegrino, & Berman, 2024).

The Role of Emerging Technologies in an IHCD-Informed R&D Infrastructure

Several scholarly, societal, and market-based trends are influencing the work of measurement professionals and the design and implementation of assessment tools and practices. Returning to Mislevy's (2019) conceptualization of assessment as "situated in social contexts, shaped by purposes, and centered on students' developing capabilities for valued activities (p. 164)," it is clear that the growing recognition of assessment's role in propagating systemic inequities must influence how we work to situate assessment in learners' diverse, multifaceted contexts. Additionally, the purposes of assessment are changing, and society's

notions regarding what activities are "valued" will inevitably evolve in response to a changing economic and civic landscape. Other chapters in this volume discuss some of these factors in detail. In this section, we focus on one potentially influential set of conditions—namely, advances in applications of technology to assessment and, in particular, the growing availability and influence of generative AI. The role of technology in an IHCD—infrastructure is likely to become more significant over time, in that most advances in assessment systems are often leveraging a more technology-enabled approach to how we assess student learning and educational outcomes. The role of technological capabilities as a way to improve our assessment paradigms will always and inevitably influence the assessment field, yet how we approach their arrival and early use is an area that will require significant improvement and a much more solid evidence base than currently exists.

An R&D infrastructure that supports IHCD-aligned practices includes a place where curiosity and early adoption can occur in a safe, cautious way. These technologies will influence our assessment paradigms, even if those actors are "external" to the organizations in which assessment R&D is their primary focus. As seen in the current market for curriculum and instructional tools, venture capital, private equity, and those in non-educational fields have exerted increasingly significant influence on the learning technologies used in schools. One question we may want to answer is how might the assessment field expand to include a more robust place within its infrastructure to hold this curiosity and faster-paced exploration in service of leading edge assessment R&D?

Responsibly Leveraging Advances in AI

The release of ChatGPT in November 2022 launched a wide-ranging debate that has raged and morphed ever since among advocates, skeptics, and everyone in between. A narrower version of this debate has been prominent within the measurement community, consuming pages of journals and books along with panel discussions at conferences. These conversations are valuable, particularly when they include representatives from diverse groups and with varying experiences and viewpoints. Given how much we don't know, and how rapidly the generative AI landscape is advancing, it is critical for assessment system designers to avoid a knee-jerk rush to "innovation" or a panicked avoidance of all things AI. Instead, an approach that combines a cautious embrace of new ideas

with an iterative, user-focused approach to testing and refining tools is likely to offer the best chance to create systems that promote the desired outcomes while limiting harm. School district leaders and other experts have been issuing guidance regarding appropriate AI use[1], though most of that guidance devotes relatively little attention to assessment. In the rest of this section, we do not attempt to lay out the many ways in which generative AI might improve or detract from high-quality assessments. Instead, we emphasize the ways in which IHCD can support responsible applications of AI to assessment systems.

Continue to Build the Evidence Base

Despite the vigorous enthusiasm that generative AI has prompted in many circles, evidence to inform decisions about how it should be incorporated into assessments is, with a few exceptions (e.g., methods for automated essay scoring), quite nascent. Because the adoption of AI for assessment is inevitable, an inclusive assessment R&D infrastructure must prioritize work to test the effects of promising AI applications on score validity, reliability, and fairness. Innovation in assessment is clearly needed, but iInnovation for the sake of innovation is unlikely to yield the intended benefits and could result in significant harms.

Ensure Stakeholder Understanding of, and Support for, AI Applications

An IHCD approach is well-suited to addressing the potential benefits and risks of AI. By engaging all relevant groups in conversations designed to enhance their understanding of AI while also eliciting their concerns and hopes regarding AI applications, developers can help ensure responsible adoption. It is especially crucial for these groups to understand the likely implications of incorporating (or not incorporating) AI for a particular assessment feature. Additionally, the increasing availability of tools to support development of "home-grown" AI tools[2], increases the urgency of engaging in collaborative conversations and reaching a collective agreement on specific AI applications.

1 For a particularly promising example, see this guidance developed by Santa Ana Unified School District: https://www.sausd.us/cms/lib/CA01000471/Centricity/Domain/6/DRAFT%20-%20SAUSDs%20AI%20Compass%20Your%20Guide%20to%20Navigating%20an%20AI%20World%209.pdf

2 https://openai.com/index/introducing-gpts/

Explore How AI Might Promote Coherence

One promising approach to pursuing the vision of coherent assessment systems described earlier is through steps that diminish the likelihood that large-scale assessments will negatively affect teaching and learning (Marion et al., 2024; Hamilton & Martinez, 2024). Applications of AI to large-scale assessments could contribute to this goal in a few ways, including through improved approaches to adaptivity that could reduce test length and through new techniques for personalizing assessments that could potentially improve alignment with other parts of the assessment system.

Consider Benefits and Drawbacks of Using AI to Promote Equity and Cultural Responsiveness

As we consider the role of the seven principles of assessment in the service of learning, advanced technical capacities have a role to play in advancing a more individualized and equitable learning environment. For example, in order to advance an assessment system that allows for individualization that is responsive to that learner's specific needs (Principles 3,5,6) or an assessment system that provides more robust and relevant results (Principle 7), the infrastructure in which it sits must be more interoperable; must be more secure; and, must allow for more advanced analytics that are combined with greater attention to data literacy support for its users. However, in order to ensure that these capabilities don't duplicate the harm that has been evidenced in the past, as discussed earlier in this chapter, this infrastructure must consider how it simultaneously advances a more inclusive and human-centered approach.

IHCD Applications to Policy, Practice, and Use: The Role of Establishing an Assessment Culture Rooted in Transformation

Systems of assessment, as previously mentioned in this chapter, are not only shaped by advances in the sciences, or the application of new technologies, they are also molded by the cultural and political environments in which they are situated. In addition, the infrastructure needed to support these advances will be more robust if it is guided and resourced at the state and federal levels. This intentional shaping of assessment paradigms, even through policy, can help to ensure that there is widespread use and application of these principles at a larger scale. While there is certainly a directionality when it comes to how policy shapes assessment practices, including the factors that support effective

implementation, by elevating a more inclusive and human-centered approach into how we craft policy, it is possible to occasion a more equitable and effective paradigm for assessment.

There are several considerations that can be advanced when thinking about the role of ICHD in how we shape the culture of assessment, and we offer a few levers for consideration. For example, how we support educator and learner use, the expansion of our ideologies within our generation and analysis of assessment data, and how we might advocate for a more inclusive approach in the design and application of policy can provide a compelling set of use cases for how we may move forward in ensuring that ICHD-aligned assessment paradigms are a fixture in the future of education.

We anticipate that the marketplace of integrated digital curriculum and assessment tools will continue its rapid growth and that educators will face an increasingly complex set of decisions regarding what tools to adopt, in what contexts, and for what purposes. Those who prepare and support educators—including classroom teachers as well as administrators who often make purchasing and adoption decisions—should develop research-based guidelines for adoption and use. Such guidelines should address questions such as how educators can make sense of the reams of data generated, how they can achieve policy objectives such as disaggregation of data and personalization of instruction to meet individual student needs, and how to interact effectively and appropriately with chatbots and search tools. Aggressive marketing of software and other tools for classroom and school use has created confusion and hindered adoption of balanced assessment (Hamilton & Martinez, 2024) and these challenges are only likely to grow.

Some Promising Efforts toward Inclusive Assessment System Design

In recent years, efforts to make sure that the R&D process, including assessment-related R&D, centers inclusion and accessibility have grown. These efforts span philanthropic organizations, such as the Chan Zuckerberg Initiative (CZI), The Spencer Foundation, and the Gates Foundation; non-profit research organizations, such as the American Institutes for Research (AIR) and the Advanced Education Research and Development Fund (AERDF); and even federal-level programs, such as National Center for Advanced Development in Education (NCADE) at the Institute for Education Sciences (IES) and the National Science Foundation (NSF). The Alliance for Learning Innovation convened a group of researchers, practitioners,

policymakers, and others to develop recommendations related to inclusive R&D; these include investing in collaborative work that honors the perspectives of diverse groups and using measurement strategies that generate more accurate and useful data for all participants (Alliance for Learning Innovation, 2024).

As a concrete example of ICHD, the Measures for Early Success Initiative, housed at the MDRC (Manpower Demonstration Research Corporation) seeks to reimagine the assessment landscape, especially the PreK assessment systems that operate at local and state levels. Through this process, they are leveraging a clear human-centered design approach, called participatory R&D, that ensures that families, educators, and learners themselves are a part of the assessment design process—even in tools that are being used at scale across the nation. This year-long process requires that development teams meet on a consistent basis with educators, caregivers, and learners, and that they weave the information gleaned from them directly into the tools they are creating.

Conclusion

As we close this chapter, we invite readers to consider what priorities might be important for the field in service of seeking an evolution to a more inclusive, human-centered set of assessment paradigms and the implications of taking on this new lens in support of a more equitable set of assessment paradigms for education. Achieving the ambitious vision described in this chapter and in the rest of the volume will require a collaborative approach that engages educators, policymakers, education support providers, measurement experts, families, and, of course, the young people who have the greatest stake in the outcomes of these efforts.

References

Ackerman, T.A., Bandalos, D.L., Briggs, D.C., Everson, H.T., Ho, A.D., Lottridge, S.M., Madison, M.J., Sinharay, S., Rodriguez, M. C., Russell, M., von Davier, A.A. and Wind, S.A. (2024). Foundational competencies in educational measurement. *Educational Measurement: Issues and Practice.* https://onlinelibrary.wiley.com/doi/10.1111/emip.12581

Alliance for Learning Innovation (2024). *Inclusive Education R&D.* https://fas.org/wp-content/uploads/2024/05/ALI-Taskforce-Brief-Inclusive-RD.pdf

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Armstrong, M., Dopp, C., & Welsh, J. (2020). *Design-based research.* https://edtechbooks.org/studentguide/design-based_research

Assessment Work Group. (2019). *Student social and emotional competence assessment: The current state of the field and a vision for its future.* Chicago, IL: Collaborative for Academic, Social, and Emotional Learning. https://measuringsel.casel.org/wp-content/uploads/2019/09/AWG-State-of-the-Field-Report_2019_DIGITAL_Final.pdf

Berg, J., Osher, D., Same, M. R., Nolan, E., Benson, D., & Jacobs, N. (2017). I*dentifying, defining, and measuring social and emotional competencies.* American Institutes for Research. https://www.air.org/sites/default/files/2021-06/Identifying-Defining-and-Measuring-Social-and-Emotional-Competencies-December-2017-rev.pdf

Burrus, S. H. Rikoon, & M. W. Brenneman. (Eds.). A*ssessing Competencies for Social and Emotional Learning: Conceptualization, Development, and Applications.* New York: Routledge. https://doi.org/10.4324/9781003102243

Cantor, P. (2021). All children thriving: A new purpose for education. *American Educator, 45*(3), 14–26. https://files.eric.ed.gov/fulltext/EJ1322300.pdf

Casillas, A., Roberts, B., & Jones, S. (2022). An integrative perspective on SEL frameworks. In J. Burrus, S. H. Rikoon, & M. W. Brenneman (Eds.), *Assessing Competencies for Social and Emotional Learning: Conceptualization, Development, and Applications* (pp. 9–27). Routledge. https://doi.org/10.4324/9781003102243-3

Darling-Hammond, L.& Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st-century standards of learning.* Stanford Center for Opportunity Policy in Education.

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8. http://www.jstor.org/stable/3699927

Domaleski, C., D'Brot, J., Pinsonneault, L., Gong, B., & Brandt, C. (2023). *The path forward for school accountability: Practical ways to improve school accountability systems now. Center for Assessment.* https://www.nciea.org/wp-content/uploads/2023/06/Path-Forward-For-School-Accountability-FINAL.pdf

Evans, C. (2024). Stop trying to use commercial interim tests for instructional purposes: A case for "strategic abandonment." *CenterLine Blog, Center for Assessment.* https://www.nciea.org/blog/strategic-abandonment-and-interim-assessment/

Faxon-Mills, S., Hamilton, L. S., Rudnick, M., & Stecher, B. M. (2013). New assessments, better instruction? Designing assessment systems to promote instructional improvement. RAND. https://www. rand.org/pubs/research_reports/RR354.html

Foster, N. and M. Piacentini (Eds.). (2023). *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Goldman, S., & Lee, C. (2024). Human learning and development: Theoretical perspectives to inform assessment systems. In Marion, S. F., Pellegrino, J. W., & Berman, A.I. (Eds.), *Implementation and use of balanced assessment systems* (pp.48–29). National Academy of Education.

Gordon, E. W., Pelligrino, J. (Eds.). (2013). Technical Report of the Gordon Commission on the Future of Assessment in Education. *Educational Testing Service.* https://www.ets.org/Media/Research/pdf/gordon_commission_technical_report.pdf

Harvard University (undated web page). *Explore SEL.* http://exploresel.gse.harvard.edu/

Hamilton, L.S., & Landl, E. (2023). Improving accountability through authentic community engagement. *CenterLine Blog, Center for Assessment* https://www.nciea.org/blog/improving-accountability-through-authentic-community-engagement/

Hamilton, L.S., & Martinez, J.F. (2024). Policy influences on ambitious classroom instruction, assessment, and learning. In Marion, S. F., Pellegrino, J. W., & Berman, A.I. (Eds.), *Implementation and use of balanced assessment systems* (pp.274–308). National Academy of Education.

Hamilton, L.S., & Murray, O. (2023). *Accelerating progress in U.S. education: Key lessons from other federal R&D investments in technology and innovation.* American Institutes for Research. https://www.air.org/resource/brief/accelerating-progress-us-education-key-lessons-other-federal-rd-investments

Hamilton, L.S., Olivera-Aguilar, M., & Rikoon, S.H. (2024). *Cultivating civic learning and engagement in U.S. schools. American Institutes for Research.* https://www.air.org/sites/default/files/2021-06/Identifying-Defining-and-Measuring-Social-and-Emotional-Competencies-December-2017-rev.pdf

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth, 17*(4), 267–321. http://www.jstor.org/stable/23324889

Jagers, R., Lozada, F. T., Rivas-Drake, D., & Guillaume, C. (2017). Classroom and school predictors for civic engagement among Black and Latino middle school youth. *Child Development, 88*(4), 1125–1138. https://doi.org/10.1111/cdev.12871

Jagers, R.J., Rivas-Drake, D., & Borowski, T. (2018). *Equity & social and emotional learning: A cultural analysis.* Collaborative for Academic, Social, and Emotional Learning. https://drc.casel.org/uploads/sites/3/2019/02/Equity-Social-and-Emotional-Learning-A-Cultural-Analysis.pdf

Koretz, D. (2008). *Measuring up: What educational testing really tells us.* Harvard University Press.

Koretz, D., & Hamilton, L.S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp.531–578). American Council on Education/Praeger.

Lampen, C. (2022, April 19). What is the conservative beef with "social-emotional learning"? *The Cut.* https://www.thecut.com/2022/04/conservative-backlash-social-emotional-learning.html

Lee, C. D., White, G., & Dong, D. (Eds.). (2021). *Educating for civic reasoning and discourse.* National Academy of Education.

Linzarini, A., & da Silva, D.C. (2024). *Innovative tools for the direct assessment of social and emotional skills.* Organisation for Economic Co-operation and Development. https://www.oecd.org/education/innovative-tools-for-the-direct-assessment-of-social-and-emotional-skills-eed9bb04-en.htm

Lovelace, T.S., McMurtry, T., & Kendall-Brooks, L. (2024). Equitable and culturally relevant social and emotional competency assessment. In J. A. Durlak, C. E. Domitrovich, R. P., & Mahoney (Eds.), *Handbook of Social and Emotional Learning: Research and Practice.* (2nd. ed.). New York: Guilford.

Marion, S. F., Pellegrino, J. W., & Berman, A.I. (Eds.), *Implementation and use of balanced assessment systems.* National Academy of Education.

McCullough, B. (2018). *How the Internet Happened: From Netscape to the iPhone.* Blue Cypress Books.

Mehta, J. (2025). Education Dept. warns schools: Eliminate DEI programs or lose funding. *NPR*. https://www.npr.org/2025/04/03/nx-s1-5350978/trump-administration-warns-schools-about-dei-programs

Mislevy, R. J. (2019). Advances in Measurement and Cognition. *The ANNALS of the American Academy of Political and Social Science, 683*(1), 164–182. https://doi.org/10.1177/0002716219843816

National Academies of Sciences, Engineering, and Medicine. (2019). *Monitoring educational equity.* National Academies Press. https://nap.nationalacademies.org/catalog/25389/monitoring-educational-equity

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* U.S. Government Printing Office.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* National Academy Press.

National Urban League & UNIDOS (2022). *Education assessment, accountability, & equity.* https://nul.org/sites/default/files/2023-04/FOAA_Final%20Phase%201%20Report_April2023_final%20.pdf

Norman, D. (2013). *The Design of Everyday Things*. Basic Books. https://doi.org/10.15358/9783800648108

Osher, D., Cantor, P., Berg, J., Steyer, L., & Rose, T. (2020). Drivers of human development: How relationships and context shape learning and development. *Applied Developmental Science, 24*(1), 6–36. https://doi.org/10.1080/10888691.2017.1398650

Rikoon, S.H., Hamilton, L.S., & Olivera-Aguilar, M. (2024). *Social and emotional learning in U.S. schools.* American Institutes for Research. https://www.air.org/sites/default/files/2024-04/social-emotional-learning-US-schools.pdf

Roschelle, J. M., Pea, R. D., Hoadley, C. M., Gordin, D. N., & Means, B. M. (2000). Changing how and what children learn in school with computer-based technologies. *The Future of Children,* 76–101. https://doi.org/10.2307/1602690

Schneider, M. (2023, February 22). Innovation in the education sciences (the new IES). *IES Blog*. https://ies.ed.gov/director/remarks/02-02-2023.asp

Schwartz, S. (2021, June 11). Map: Where critical race theory is under attack. *Education Week.* https://www.edweek.org/policy-politics/map-where-critical-race-theory-is-under-attack/2021/06

Smith, M., & O'Day, J. (1991). Systematic school reform. In S. Fuhrman and B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233–268). Falmer Press.

Vinnakota, R. (2019). *From civic education to a civic learning ecosystem: A landscape analysis and case for collaboration.* Red & Blue Works. https://rbw.civic-learning.org/wp-content/uploads/2019/12/CE_online.pdf

Zaber, M. A., Karoly, L. A., & Whipkey, K. (2019). *Reimagining the workforce development and employment system for the 21st-century and beyond.* RAND Corporation. https://doi.org/10.7249/RR2768

Designing and Developing Educational Assessments for Contemporary Needs

Kristen Huff

Abstract

Conventional assessment design and development approaches that have served the field for decades are struggling to meet today's growing educational needs. To adequately handle the complexity and scope of the knowledge we aim to measure, assessments must be designed with as much rigor and clarity as possible. The Principled Assessment Design approach provides solutions to these challenges that remain coherent across all elements of an assessment system. The three iterative phases of PAD foster a deeper shared understanding of student cognitive processes and the role of performance level descriptors. Rather than reactive, ad hoc validation based on analysis of empirical data, a strong inferential, evidentiary, and validation argument begins to take shape proactively in the design process. The documentation generated throughout each iteration provides design tools that inform development of PLDs and task models. PAD therefore offers an adaptable framework to address evolving educational goals while embracing our growing understanding of cultural responsiveness. The principles of PAD are essential in creating fair and accessible assessments while maintaining the integrity of the constructs being measured. The assessment community faces an inflection point amid complex and contemporary demands, expectations, and capabilities; proponents of PAD are poised to meet those needs.

Educational assessment stands at a critical inflection point. Conventional assessment design and development approaches that have served the field for decades are being challenged by increasing demands for assessments that are more authentic, informative, actionable, engaging, and accessible for all students. Principled Assessment Design (PAD) is an alternative approach that addresses the challenge of creating culturally and linguistically responsive, yet fair and well-designed, assessments. Using the three iterative phases of PAD assessment designers can address these goals from the outset with strong inferential, evidential, and validation arguments. Through a focus on student cognition, consistent documentation, and continuous re-evaluation, this knowledge base can be built upon to evolve with our understanding of fairness in assessment.

The Case for Principled Assessment Design

Historically, large-scale assessments were designed primarily to rank order students, and the primary, and perhaps only, interpretation from the resulting scores was the percentile rank of the test taker in comparison to a national norm-referenced distribution (National Research Council, 2001; Shepard, 2000). The objective was not to support claims about what a student knows and can do in the tested domains. Although many assessments are now called upon to provide strong inferences about what students have learned and what they need to learn to reach a particular performance level, assessment design and development practices are still largely rooted in the norm-referenced paradigm.

There are at least three reasons that conventional approaches to assessment design and development are insufficient for meeting today's educational needs. First, there is a growing complexity of what we aim to measure, such as mathematical practices, three-dimensional science learning, and collaborative problem-solving (NGA Center & CCSSO, 2010; NGSS Lead States, 2013). Second, users are demanding that K–12 assessments serve multiple purposes. Rather than add to the proliferation of testing that occurs within any given school year, it is incumbent upon the industry to design tests from the outset that have clear, strong validation arguments that can be built upon for additional use cases (AERA, APA, & NCME, 2014; Hart et al., 2015; Huff & Goodman, 2007). Third, assessment quality is under more scrutiny than ever (OESE, 2018). Our industry needs to do a better job at ensuring that assessments that are designed to make claims about student learning are designed with as much rigor and clarity as possible to do just that.

Principled Assessment Design (PAD) is an approach to assessment design that offers solutions to these contemporary challenges. PAD is a set of practices and documentation that helps ensure coherence across all elements of an assessment system. The provenance of PAD is rich, including but not limited to construct-centered measurement (Messick, 1994; Wilson, 2005), cognitive design systems (Embretson, 1998; Rupp & Leighton, 2017), evidence-centered design (Huff et al., 2010; Mislevy, 2006; Mislevy et al., 2003; Mislevy & Haertel, 2006; Pearlman 2008a, 2008b), principled design for efficacy (Nichols et al., 2016) and assessment engineering (Luecht, 2013).

There are at least three characteristics of PAD that distinguish it from conventional assessment design. The first is the role of learning science (Pellegrino et al., 2016). In PAD, where the primary purpose of the assessment is to measure where students are along a learning trajectory, the construct must be defined and all decisions about assessment design must be rooted in the science of how students learn and build knowledge. The second is the practice of documentation that can serve as design tools throughout assessment development and assessment interpretation (and if needed, even beyond, to curricular, instructional, or teacher professional learning materials). The third is a mindset of inquiry, where reasoning from imperfect evidence requires continuous interrogation of assumptions about the inferential, evidential, and validation arguments.

There are three essential phases in PAD. These phases are iterative rather than linear. The first is analyzing the domain with respect to the science of how students learn and build knowledge (Ewing et al., 2010; Pellegrino et al., 2016). In this phase, the construct and the targets of measurement for the assessment are defined. The next phase is modeling the domain, where the approach to cognition for the assessment is defined, as well as the performance level descriptors (PLDs). The final phase of design is to create design patterns for the tasks (items) which requires a shared understanding of the difficulty drivers for the tasks (Luecht, 2019). The difficulty drivers are directly informed by the approach to cognition that is defined in the second phase.

Just as there is iteration between the three general phases of PAD, there is not necessarily a clear distinction of when design ends and development begins. The point is that there needs to be a design phase to develop a shared understanding of

all of the above—and the appropriate documentation—rather than simply articulating the test specifications and starting item development, which is common practice in conventional test development (Schmeiser & Welch, 2006).

The Cognitive Approach in PAD

As mentioned above, one of the distinguishing characteristics of PAD is the explicit grounding in learning science, and as a result, a deeper focus on the role of cognition in both learning and the assessment of learning. A shared understanding—and clear documentation—of how student cognitive processes will be treated in PLDs and therefore in the assessment is a key component of the PAD.

When defining the cognitive approach that will be used to develop PLDs and therefore tasks, the interdisciplinary assessment design team investigates the answers to questions like the following, develops a shared understanding of the answers, and documents the answers in a way that can be used as design tools in the development of PLDs and, later, task models:

1. What does learning science say about how student thinking evolves from novice to proficient in this domain?
2. What constitutes observable evidence of that thinking?
3. How rigorous should proficiency be for this grade level?
4. How much learning should be expected in a given grade level for this domain?
5. What skills should be taught and assessed in this grade level to support the student learning journey?
6. What is observable evidence of each skill at each novice, proficient, advanced for this grade level?

Engaging in this line of inquiry among the assessment design team has many benefits, not the least of which is a strong inferential, evidentiary, and validation argument that is defined starting with design rather than the way validation typically occurs in conventional assessment design: post-hoc and based not solely but mostly on analyses of empirical data.

Performance Level Descriptors: The Backbone of Assessments Designed to Measure Learning

PLDs play a critical role in ensuring that the assessment is designed to support inferences about where students are along a latent proficiency continuum. In conventional approaches to assessment development, PLDs were developed after the assessment was developed as an input into the standard-setting process. In PAD, PLDs embody the approach to cognition and are the foundation of task design, informing the desired psychometric properties of the scale, and score interpretation.

Addressing Key Challenges in Assessment Design

Accessibility and Cultural Responsiveness

Assessment designers face the dual challenge of creating assessments that are accessible to all students while maintaining the integrity of the constructs being measured. For example, audio options that read text aloud may be essential accommodations for some students but could fundamentally alter what's being measured in a reading comprehension assessment. In these cases, we argue that assessment designers must have clear, strong inferential, evidential, and validation arguments so that they can nimbly engage in discussions about what inferences about student learning can and cannot be supported with various accessibility features, and whether a redefinition of the target of measurement is warranted.

Similarly, our understanding of fairness in assessment is evolving beyond merely avoiding bias to actively embracing cultural and linguistic responsiveness. Rather than simply stripping items of cultural context, there is growing recognition that assessments should reflect and value the diversity of students' lived experiences and cultural backgrounds. This might mean including garden contexts that represent urban community gardens and rural farms rather than exclusively suburban backyards, or ensuring that historical passages don't erase the experiences of marginalized communities.

PAD provides the structured framework needed to thoughtfully navigate these tensions. By clearly articulating the intended targets of measurement and assumptions about what constitutes construct-relevant versus construct-irrelevant variance, assessment designers can make principled decisions about accessibility features, accommodations, and cultural representation (CAST, 2018; Solano-Flores, 2019; World Wide Web Consortium, 2018).

Student Engagement and Motivation

The role of student engagement and motivation in assessment performance has gained increased attention, particularly for assessments that lack direct consequences for students, such as interim or embedded assessments. If students aren't engaged nor motivated to perform at their best, assessment results likely underestimate their learning (Tsai et al., 2020; Wise & DeMars 2005).

This challenge highlights the importance of integrating User Experience (UX) design expertise into assessment development. UX designers bring crucial perspectives on creating assessment experiences that are intuitive, transparent, and even enjoyable. They help ensure that navigation systems, visual elements, and interactions don't introduce construct-irrelevant barriers to performance.

Key questions that UX designers help assessment teams address include:

- Are interactions clear, easy to use, and age-appropriate?
- Is visual content accessible, equitable, and unambiguous?
- Does the experience feel familiar and consistent across items?
- Are there distracting elements that might interfere with performance?

Within a PAD framework, UX considerations become integral to task design rather than superficial enhancements. When design patterns explicitly address engagement factors and cognitive load considerations, the resulting assessments are more likely to elicit compelling evidence of where students are along their learning journey.

Looking Ahead: The Inflection Point for Assessment Design

The assessment field may be approaching what business strategists call a "strategic inflection point"—a moment when fundamental assumptions are challenged and business models are upended (Christensen, 1997; Grove, 1996). Several indicators suggest this inflection point may be imminent:

- The constructs we seek to measure are becoming increasingly complex
- Educators and policymakers are dissatisfied with both the quantity and quality of current assessments
- There are growing demands for assessments to serve multiple purposes simultaneously

- Technology is creating new possibilities for assessment design and delivery—especially AI
- Learning science research is advancing at a rapid pace

As Reed Hastings of Netflix discovered with streaming video, the timing of inflection points is difficult to predict (Hastings & Meyer, 2020). However, assessment designers who embrace principled approaches now will be well-positioned when the industry reaches its tipping point.

The adoption of PAD, with its emphasis on cognitive foundations and explicit design rationales, represents a significant shift from conventional assessment development. Like Hastings waiting for streaming to take off, proponents of PAD have been anticipating its widespread adoption for over two decades. The convergence of complex measurement demands, technological capabilities, and evolving stakeholder expectations may finally create the conditions for PAD to become standard practice rather than the exception.

Conclusion

The future of educational assessment lies in the thoughtful integration of principled design approaches, cultural responsiveness, engaging user experiences, and emerging technologies. By building assessments that provide insights, deliver meaningful information, and cohere with instruction, we can fulfill the promise of assessment as a tool that genuinely supports teaching and learning.

The assessment community faces a choice: continue with conventional approaches that have served adequately in the past or embrace more rigorous, transparent methods that can meet complex, contemporary demands. PAD, with its explicit cognitive models and carefully constructed PLDs, offers a framework not just for better assessments, but for educational experiences that truly reveal what students know and can do—and point the way toward their continued growth.

As the complexity of educational goals increases and the technologies available to measure them evolve, the principles of PAD become not just beneficial but essential to creating assessments worthy of the time students and educators invest in them. The field may be approaching its inflection point; the question is whether we will be ready when it arrives.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.*

CAST. (2018). *Universal Design for Learning Guidelines, version 2.2.*

Christensen, C. M. (1997). *The innovator's dilemma.* Harvard Business School Press.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods,* 3, 300–396.

Ewing, M., Packman, S., Hamen, C., & Thurber, A. C. (2010). Representing targets of measurement within evidence-centered design. *Applied Measurement in Education, 23*(4), 325–341.

Grove, A. S. (1996). *Only the paranoid survive: How to exploit the crisis points that challenge every company.* Doubleday.

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., & Spurgeon, L. (2015). *Student testing in America's great city schools: An inventory and preliminary analysis.* Council of the Great City Schools. http://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Testing%20Report.pdf

Hastings, R., & Meyer, E. (2020). *No rules rules: Netflix and the culture of reinvention.* Penguin.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19–60). Cambridge University Press.

Huff, K., Nichols, P., & Schneider, C. (in press). Designing and Developing Educational Assessments. In Linda L. Cook and Mary J. Pitoniak (Eds.), *Educational Measurement.* Oxford University Press.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education,* 23, 310–324.

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology, 14*(1), 1–38.

Luecht, R. M. (2019, January). *Strengthening Claims-Based Interpretations and Uses of Local and Large-scale Science Assessment Scores (SCILLSS): The role of performance level descriptors for establishing meaningful and useful reporting scales in a principled design approach* [White paper]. Nebraska Department of Education. https://www.scillsspartners.org/wpcontent/uploads/2019/02/SCILLSS_PLD_WhitePaper_V1812-02_FINAL_2_7_19.pdf

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13–23.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). American Council on Education.

Mislevy, R. J., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25,* 6–20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–67.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards.* http://www.corestandards.org

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* The National Academies Press. https://doi.org/10.17226/10019

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states.* The National Academic Press. https://www.nextgenscience.org/search-standards

Nichols, P., Ferrara, S., & Lai, E. (2016). Principled design for efficacy: Design and development for the next generation of assessments. In H. Jiao & R. W. Lissitz (Eds.), *The next generation of testing: Common Core Standards, Smarter Balanced, PARCC, and the nationwide testing movement* (pp. 49–81). Information Age Publishing.

Office of Elementary and Secondary Education. (2018, September 24). *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process.* U.S. Department of Education. https://www.ed.gov/sites/ed/files/2023/11/assessmentpeerreview.pdf

Pearlman, M. (2008a). Chapter 3: The design architecture of NBPTS certification assessments. In R. E. Stake, S. Kushner, L. Ingvarson, & J. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards: Advances in program evaluation* (Vol. 11, pp. 55–91). Emerald Group Publishing.

Pearlman, M. (2008b). Finalizing the test blueprint. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 227–258). Routledge.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81.

Rupp, A. A., & Leighton, J. P. (Eds.). (2017). *The handbook of cognition and assessment: Frameworks, methodologies, and applications.* Wiley–Blackwell.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). National Council on Measurement in Education and American Council on Education.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation. *Frontiers in Education, 4,* 43. https://doi.org/10.3389/feduc.2019.00043

Tsai, Y.-S., Whitelock-Wainwright, A., Chiu, Y.-L., He, Y., & Gašević, D. (2020). User experience design for technology-enhanced learning: A systematic review. *British Journal of Educational Technology, 51*(6), 2005–2033.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Lawrence Erlbaum Associates.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and possible solutions. *Educational Assessment, 10*(1), 1–17.

World Wide Web Consortium. (2018). *Web Content Accessibility Guidelines (WCAG) 2.1.* https://www.w3.org/TR/WCAG21/

VOLUME I | SECTION 2

Evidence and Research Bases that Inform the Design for Assessment in the Service of Teaching and Learning

From *How People Learn* to Assessment that Serves Learning: Toward an Assessment Ecosystem Grounded in the Sciences of Teaching and Learning

Eleanor Armour-Thomas and Eric M. Tucker

Building on Gordon's Series Introduction (Gordon, 2025), Section 2 advances assessment as integral to learning, drawing on learning and developmental sciences, sociocognitive frameworks, and measurement methodologies (Baker & Gordon, 2014; Goldman & Lee, 2024). The chapters argue that assessment should be grounded in how people learn and developed and designed to inform instruction and learning (National Research Council, 2001; Pellegrino, 2025; Pellegrino, DiBello, & Goldman, 2016; Pea, Lee, Nasir, & McKinney de Royston, 2025).

Re-grounding Assessment in the Science of Learning and Development

Decades of research identify learning as a complex, relational process shaped by interactions among learners and their social, cultural, and physical contexts (Cantor, Osher, Berg, Steyer, & Rose 2019; Cantor, Gomperts, Lerner, Pittman, & Chase, 2021; Cantor & Osher, 2021). Learning is inseparable from context—it is inherently social, emotional, and cultural, shaped by all aspects of a person's development and everyday life.

Applying this dynamic systems perspective, Lerner and Cantor (2025) argue that assessment must account for the "individual-in-context" by capturing the interplay between a learner and their environment over time. Such an approach requires that we measure not only learners' knowledge but also the "nurturant" supports in their lives—the relationships and contextual factors that enable learning (Lerner & Cantor, 2025).

Pea, Nasir, Lee, and McKinney de Royston (2025) argue that assessments aimed at improving learning must start with how learning actually unfolds, centering learners' social and cultural assets and contexts. Assessments should be grounded in learners' real-world experiences, not context-free tasks. Further, the design of assessment tasks should draw on everyday and community language and cultural practices that value human variation (Gordon, 1995).

Badrinarayan, Darling-Hammond, and Bennett (2025), in their discussion of socioculturally responsive assessment, underscore that making learning visible is *inherently a cultural activity*. One of their key points is that *validity itself* should be reconceived to include cultural validity (Gordon Commission, 2013; Mislevy, 2018). Thus, human variation and the science of learning call on us to expand our conception of what assessment quality means, incorporating efficacy and responsiveness as priorities among principles that guide the design and use of learning-focused assessments (Baker, Everson, and Tucker, 2025).

Integrating Assessment with Pedagogy to Support Learning

Armour-Thomas (2025) argues that to serve learning, assessment must be a core pillar of pedagogy, mutually reinforcing curriculum and instruction (Shepard, 2019; Armour-Thomas, 2025). Assessment should thus elicit and illuminate student thinking—not just outcomes—and drive continuous feedback cycles (Gordon Commission, 2013; Ruiz-Primo & Furtak, 2024).

Building on this, Armour-Thomas, Darvin, and Hughes (2025) argue that assessment should be tailored to the unique ways of reasoning and building knowledge within each discipline. Assessment should thus be domain-specific and aligned with each discipline's pedagogical content knowledge.

Hattie, Sireci, and Baker (2025) argue that improvement hinges on shifting mindsets so assessment is increasingly used to diagnose learning, understand student thinking and strategies, and inform next steps. Creating an error-tolerant classroom climate ensures assessment is embraced as an opportunity rather than feared, allowing mistakes to be seen as informative.

Principles for Instructionally Impactful and Effective Assessment Systems

Badrinarayan (2025) argues that large-scale assessments must be reimagined to support teaching and learning: authentic, curriculum-anchored, educative, developmental, culturally responsive, and timely. Collectively, these principles—illustrated through practical examples like states piloting performance-based assessments—underline that redesigning assessments to better serve teaching and learning is both achievable and necessary.

Marion and Evans (2025) share criteria for instructionally useful assessments: cognitive complexity, close alignment to the curriculum, timely and fine-grained results, and the use of authentic, open-ended tasks. This analysis resonates with the call for *balanced assessment systems* that privileges rich classroom learning environments (Marion, Pellegrino, and Berman, 2024).

LeMahieu and Cobb (2025) introduce *practical measurement for improvement*, an approach embedding measures within daily practice to support real-time learning and continuous improvement. Rooted in improvement science, practical measurement emphasizes frequent, fine-grained assessments specifically designed to inform instructional refinement, rather than acting as endpoint audits. These measures are timely, context-specific, and purpose-built to yield actionable insights aligned with clear improvement goals. By explicitly asking "what works, for whom, under what conditions"—this approach helps educators identify effective practices.

Section 2 offers a research-grounded, forward-looking case for assessment that serves learning—especially when embedded in feedback-rich pedagogy (LeMahieu & Cobb, 2025). Furthermore, it clarifies that learner-centered assessment must reflect the variations of how learners learn as well as the conditions and supports necessary and sufficient for deep learning and its improvement over time.

References

Armour-Thomas, E. (2025). Dynamic pedagogy: A perspective for integrating curriculum, instruction, and assessment in the service of learning at the classroom level. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Armour-Thomas, E., Darvin, J., & Hughes, G. B. (2025). Assessment as a pillar of pedagogy in support of learning in AP Research and mathematics education courses. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Badrinarayan, A. (2025). Reimagining state assessments in service of teaching and learning: Design principles for instructionally relevant assessments. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Badrinarayan, A., Bennett, R. E., & Darling-Hammond, L. (2025). Perspectives on socioculturally responsive assessment in large-scale systems. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Baker, E. L., & Gordon, E. W. (2014). From the assessment of education to the assessment for education: Policy and futures. *Teachers College Record, 116,* 1–24.

Cantor, P., & Osher, D. (Eds.). (2021). *The science of learning and development: Enhancing the lives of all young people.* Routledge.

Cantor, P., Gomperts, N., Lerner, R. M., Pittman, K., & Chase, P. (2021). *Whole-child development, learning, and thriving: A dynamic systems approach.* Cambridge University Press.

Cantor, P., Osher, D., Berg, J., Steyer, L., & Rose, T. (2019). Malleability, plasticity, and individuality: How children learn and develop in context. *Applied Developmental Science, 23*(4), 307–337. https://doi.org/10.1080/10888691.2017.1398649

Goldman, S. R., & Lee, C. D. (2024). Human learning and development: Theoretical perspectives to inform assessment systems. In S. F. Marion, J. W. Pellegrino, & A. I. Berman (Eds.), *Reimagining balanced assessment systems* (pp. 48–92). National Academy of Education.

Gordon, E. W. (1996). Can performance-based assessments contribute to the achievement of educational equity? In D. Boykoff, B. Palmer Wolf, & D. Palmer Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities.* National Society for the Study of Education.

Gordon, E. W. (2020). *Toward assessment in the service of learning. Educational Measurement: Issues and Practice, 39*(3), 72–78.

Gordon, E. W. (2025). Series introduction: Toward assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Gordon, E. W., & Rajagopalan, K. (2016). New approaches to assessment that move in the right direction. In *The Testing and Learning Revolution: The Future of Assessment in Education* (pp. 107–146). New York: Palgrave Macmillan US.

Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment: Technical Report.* Educational Testing Service.

Hattie, J., Sireci, S. G., & Baker, E. L. (2025). Mind frames for improving educational assessment. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

LeMahieu, P., & Cobb, P. (2025). Practical Measurement for Improvement: Foundations, Design, Rigor. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I*: Foundations for Assessment in the Service of Learning. University of Massachusetts Amherst Libraries.

Lerner, R. M., & Cantor, P. (2025). Implications of a dynamic, relational-developmental-systems perspective for research design, measurement, and data analysis in the service of understanding and enhancing youth development and learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Marion, S. F., & Evans, C. M. (2025). Conceptualizing and evaluating instructionally useful assessments. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Marion, S. F., Pellegrino, J. W., & Berman, A. I. (Eds.). (2024). *Reimagining balanced assessment systems.* National Academy of Education.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement.* Routledge.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* National Academies Press. https://doi.org/10.17226/10019

Pea, R., Lee, C., Nasir, N., & McKinney de Royston, M. (2025). The cultural foundations of learning: Design considerations for measurement and assessment. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Pellegrino, J. W. (2025). Arguments in support of innovating assessments. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550

Ruiz-Primo, M. A., & Furtak, E. M. (2024). *Classroom activity systems to support ambitious teaching and assessment.* In S. F. Marion, J. W. Pellegrino, & A. I. Berman (Eds.), *Reimagining balanced assessment systems* (pp. 93–131). National Academy of Education.

Shepard, L. A. (2019). *Classroom assessment to support teaching and learning. The ANNALS of the American Academy of Political and Social Science, 683*(1), 183–200. https://doi.org/10.1177/0002716219843818

The Cultural Foundations of Learning: Design Considerations for Measurement and Assessment

Roy Pea, Carol D. Lee, Na'ilah Nasir, and Maxine McKinney de Royston

Abstract

In this chapter, we explore the implications of this key insight for the design of assessments. We begin the chapter with an exploration of insights from the most recent science on what learning is and how it happens. We draw from the Science of Learning and Development (SOLD), to explicate key ideas about the nature of learning and the kinds of learning that assessment should serve (Darling-Hammond, Flook, Cook-Harvey, Barron, & Osher, 2020). Then we turn to a discussion of assessment, underscoring that our current system of assessment in the U.S. primarily focuses on sorting, rather than learning (Goldman & Lee, 2024), and in doing so, such assessments too often reify racial and class-based disparities. We then examine how we should be thinking about assessment practices, exploring what might be optimal if we were seeking to assess deep learning. We conclude with a discussion of further implications of this perspective for the development of assessments, including the role of AI, and assessment within disciplines.

At its core, the aim of assessment is to productively understand what and how students learn, with an eye towards improving instruction to support further learning processes and enhance learning outcomes. We argue in this chapter that such an endeavor can only effectively occur if we: 1. Start from what we know about learning; 2. Take as central the social and cultural contexts of young people's lives and learning; and 3. Get clear about the purposes of our assessments.

It is a particularly rich time in education for these discussions (Baird, Andrich, Hopfenbeck, & Stobart, 2017; Baroody & Pellegrino, 2023). All indicators suggest that we have moved past the standards-based movement, with its overemphasis on summative high-stakes assessments and punitive approach to improving instruction and learning outcomes (Darling-Hammond, 2017; Kirst, 2024; Volante, Klinger, & DeLuca, 2024). However, while it is clear what has not worked, it may be less clear what might work for developing assessment and measurement systems rooted in robust theories of learning that can provide insights into the complexities of young people's thinking and learning. It is also unclear what kinds of such systems are feasible given the vast number of constraints in education systems. In this chapter, we explore the implications of taking seriously what we know about learning as we consider the design of assessment and measurement systems.

We begin the chapter with an exploration of insights from the most recent science on what learning is and how it happens. We draw from the Science of Learning and Development (SOLD)—an interdisciplinary science converging from research in the cognitive sciences, psychology, neurosciences, sociology, and other fields—which elucidates the nature of learning in ways transcending centuries-old empiricist notions of learning as a passive and culturally neutral process principally involving individual cognition (Nasir, Lee, Pea, & McKinney de Royston, 2020). We seek to ground the insights in this chapter in some key ideas from this literature about the nature of learning and the kinds of learning that assessment should serve (Darling-Hammond, Flook, Cook-Harvey, Barron, & Osher, 2020). Then we turn to a discussion of assessment, underscoring how our current system of assessment in the U.S. primarily focuses on sorting rather than learning (Goldman & Lee, 2024), and that, in doing so, it too often reifies racial and class-based disparities. The next section takes up the questions of how we should be thinking about assessment practices, exploring what might be possible and/or optimal if we sought to assess deep learning rather than focusing principally on sorting. We turn then to

a discussion of further implications of this perspective for the development of assessments, including the role of AI, and assessment within disciplines.

This chapter highlights several of the design principles for assessment that guide this volume. Specifically, it takes up four of these principles. Our arguments take as core that **assessment equity requires fairness in understanding tasks and adaptation to permit the use of different background knowledge and experience.** Our approach begins with the goal of providing equitable access to high quality teaching and learning and views it necessary for the range of students' background knowledge and experiences, from their homes, communities, and cultures, to be valued for learning to take place. We also call attention to the fact that **assessment design supports the learner's processes, motivation, attention, engagement, effort, and metacognition (self-regulation).** We know from the science of learning and development that to assess learning accurately and with integrity, the assessments themselves need to be designed in ways that works with, rather than against, the ways that learners learn. Thus, considering motivation, engagement, and the learners processes is fundamental to assessing learning well. We also argue that assessments need to be more integrated with learning, which includes the idea that at best, **assessments model the structure of expectations and desired learning over time.** They are learning opportunities in and of themselves, which must reflect the values and the learning process. And finally, our chapter underscores that **feedback from assessment results for learners, teachers, administrators, and families clearly addresses decisions and next steps.** When assessments are more aligned with the learning young people are engaging across the contexts of their lives and reflects the desires of learners themselves and their teachers, families, and communities, then the learning they support will be more effective. But that is not enough—they also must be in communication with those stakeholders to ensure clarity about decisions, implications, and next steps.

What We Know About Learning

A long-time popular conception shared by many educational researchers is that learning is something that happens primarily in the head, involving 'exposure to' and 'uptake of' facts and information. Indeed, the cognitive revolution advanced this perspective by documenting how humans use a variety of active cognitive processes in the accumulation of knowledge and for reasoning about how the world works. This perspective was instantiated in behaviorism, cognitive

processing, and cognitive development models of learning (Greeno, 1998). But the most recent science suggests that learning is a much richer and much more complex process, involving not only cognition, but also processes and systems involving emotion, identity, self-perception, and cultural context (Lee, 2024; Nasir, 2024; Nasir et al., 2020; Osher et al., 2016). Further, these systems that have been conceived of as discrete actually interweave throughout development and are rooted in evolutionary drives that make humans fundamentally social (Cantor & Osher, 2021; Immordino-Yang, 2016; Packer & Cole, 2020). One synthesis of this vast body of interdisciplinary scholarship is provided by the RISE Principles (McKinney de Royston et al., 2020; Nasir et al., 2020).

The RISE Principles offer one effort towards a theory of learning that honors the complex and multi-dimensional nature of learning. The four RISE Principles are that learning is:

1. Rooted in the evolutionary, biological and neurological systems of our bodies and minds, and inseparable from our social and cultural activities;
2. Integrated with all other aspects of development, including cognition, emotion, and the formation of identity—to establish a wide-angle view of the whole child;
3. Shaped by everyday life cultural activities, both in and out of school and across the lifespan; and
4. Experienced in our bodies through coordination with social others and the natural and designed worlds.

The first principle, that learning is *Rooted in the evolutionary, biological and neurological systems of our bodies and minds, and inseparable from our social and cultural activities,* begins at the beginning, with the very nature of our evolutionary biological, and neurological systems. It underscores how humans are fundamentally designed to work in context, in proximity to social others, and to meet our human needs for connection (Immordino-Yang, 2016; Lee, Meltzoff, & Kuhl, 2020; Packer & Cole, 2020). This principle calls our attention to the way that we are 'hard wired' for connection and social interaction. Additionally, while evolution has typically been thought of as the changes in biological systems adapting to circumstances and context overtime, social and cultural systems themselves also adapt in important ways (Packer & Cole, 2020; Turner, 2020). This

intertwining of cultural and biological systems is a core principle of development, and is played out across domains, such as brain development and emotional development (Immordino-Yang, 2016; Lee, Meltzoff, & Kuhl, 2020). Thus, to be human is to learn, because being human requires adaptability in the face of changing contexts.

The second principle, that learning is *Integrated with all other aspects of development, including cognition, emotion, and the formation of identity—to establish a wide-angle view of the whole child,* focuses on the ways in which learning involves integration across developmental domains in a whole person perspective which highlights how emotion, identity, self-perception, and cognition are all brought to bear in the learning process. Further, these processes themselves take place in the specificities of particular social, cultural, and historical contexts, which are in interaction with developmental processes such that these settings offer certain possibilities for development while creating challenges with others. It is through such interactive, situated cultural practices that social, racial/ethnic and gender identities exert influence on our developmental and learning trajectories; as histories of racism and other forms of exclusion challenge some possibilities, including the learners' social conditions, such as poorly resourced schools and fewer opportunities to learners (Darling-Hammond, 2010; Moss et al., 2008). Neuroscience research demonstrates that learning is more effective when learners feel safe and a sense of belonging (Darling-Hammond, 2023, Immordino-Yang, 2016; Steele & Cohn-Vargas, 2013), thereby providing yet another example of how the emotional aspects and social contexts are consequential for learning and integral to learning processes. In this respect, it is worth remembering that Benjamin Bloom's influential taxonomy of psychological domains for education (Bloom & Krathwohl, 1956) and its revised version (Anderson & Krathwohl, 2001) encompassed not only the cognitive and psychomotor manual/physical skills but the affective domain of growth in feelings or emotional areas. It did not treat cognitive developmental processes as dissociable nor in isolation from affective developmental ones.

The third principle, that learning is *Shaped by everyday life cultural activities, both in and out of school and across the lifespan,* focuses squarely on how social context shapes learning. It also emphasizes a point that principle one implied—that learning is ubiquitous, happening all of the time and everywhere, and is thus life-wide and life-deep (Banks, Au, Ball, Bell, Gordon, Gutiérrez, Brice-Heath, Lee, Mahiri, Nasir,

Valdés, & Zhou, 2007). Research has shown that rich and deep learning happens in a variety of settings, from families (Nasir, McKinney de Royston, Barron, Bell, Pea, Stevens, Goldman, 2020; Stevens, 2020), to refugee camps (Brice-Heath, Bellino, & Winn, 2020; Dryden-Peterson, 2016), to games (Pinkard et al., 2017), new media communications (Barron et al., 2013), to local corner stores and basketball courts (Nasir, 2000; Taylor, 2009). Not only does learning happen effectively in a range of settings, but it is noteworthy and even problematic that we tend to privilege the learning that happens in schools and thus undervalue the important knowledge young people bring to schools from their homes and communities (Barron, 2006). Learning also occurs on multiple time scales, and shifts depending on where learners are in the life course (Lee, 2024).

And finally, the fourth principle is that learning is *Experienced in our bodies through coordination with social others and the natural and designed worlds.* This principle highlights that learning involves multiple aspects of ourselves, including our physical selves. Learning, like cognition, occurs throughout our bodies, not only in the brain. Embodiment is central to learning—we learn in and through our bodies (Alibali, 2025; Kontra, Goldin-Meadow, & Beilock, 2012; Nathan, 2021; Shapiro & Spaulding, 2024). This embodiment includes touch or sensorimotor interactions, simulation processes, and kinesthetics, as well as embodied activities more directly connected to communication and expression such as gesture and dance (Vogelstein et al., 2019). The principle of learning as embodied honors that not only is human learning experienced through our bodies, but it does this through social coordinated interactions with others (McDermott & Pea, 2020) and with artifacts (Cole, 1996) created by humans. Bahktin (1981) has argued that even when we are alone, our thinking takes up ideas, beliefs and artifacts created by other human beings. Social forms like language mediate learning—i.e., we learn in and through language—as with tools like symbolic systems and computing and communication technologies (Flores, 2020; Peas, 1994; Rosa, 2016; Vakil, 2024; Valdés, 2004). Symbolic systems embody what Cole calls conceptual artifacts. Indeed, language is a mediator of learning and language is used to position learners into particular identities, whether as members of particular communities and/or as experts or novices in an activity (Green, et. al., 2020). Language is also politicized, wherein some languages or speakers—and their bodies—become stigmatized and get leveraged as rationales for denying them access to learning opportunities. This suggests that our accounts of learning also need to attend to the embodied and

physical aspects of learning, including how we learn kinesthetically and through gesture and other forms of bodily engagement (e.g., Abrahamson & Lindgren, 2014).

Overall, the RISE principles offer a way to view learning that honors learning as an expansive and multi-dimensional process; they also make clear how social and cultural contexts are central to them. If learning is this complex and multi-dimensional and involves so many different interacting and inter-connected developmental domains—what does this mean for assessment? How might we design assessments in the service of learning? We will turn to these questions shortly—first, we consider what assessment is, and what we might be trying to accomplish with it.

What Is Assessment?

Assessments in education tend to focus on academic outcomes in literacy and mathematics, in a way that privileges content knowledge, rather than complex thinking processes (Lee, 2024). The past three decades, in particular, have been an era of standards-based assessments, where students take high stakes tests in math and literacy annually, and the results are used to rank schools, determine student proficiency, and to determine teacher merit pay and school status (Kirst, 2024; Darling-Hammond et al., 2017; Nichols & Berliner, 2007). Research has shown that high stakes assessment creates conditions that undermine learning, such as teachers teaching to the test, cheating, and reducing time in subject matters that are not tested, like science and civics (Darling-Hammond et al, 2017). Further, such assessments reinforce disparities by race and social class and fail to include important knowledge that students bring from their families and communities (Goldman & Lee, 2024). They are aligned with the "dominance of restrictive conceptions of what counts as knowledge in the disciplines and the ontology of the disciplines that are currently restricted to Eurocentric histories of the disciplines." (Lee, 2024, pg. 4). In other words, the assessments we most value in the U.S. fail to capture the breadth of knowledge that constitutes learning. This is, in part, because most assessments used in the U.S. are static, one-time summative measures that are not useful for formative purposes nor to inform shifts in instruction.

This contrast between summative and formative assessment raises a key aspect of the challenge. In his influential paper, philosopher of science Michael Scriven (1967) did not consider formative and summative assessment as two different *types* of evaluation—a frequent misunderstanding. Scriven viewed them as two

different *roles* that evaluation can play. In the formative role for assessments, the teacher or evaluator is taking a constructive approach by emphasizing the input that will help improve a program of instruction and thus improve learning. Whereas in the summative role, the teacher or evaluator is determining the worth of the instructional program by understanding the quality of the learning.

When we are not clear about what assessments are for, and when we are unclear about what they do and do not allow us to understand, we are likely to design assessments that may be practically feasible but conceptually unaligned with our goals. Likewise, when assessments focus on outcomes of learning, rather than including the processes of learning, they are less useful to guide teaching and fall short of being a teaching tool for improving the learning of students.

We know that in their purpose and content, existing standardized tests can be conceptualized as cultural artifacts that are products of our nation's dominant common culture in that they are shaped by it and used for its benefit (Greenfield, 1997; Solano-Flores, 2019). Thus, another central challenge is the way that current forms of assessment, and our cultural assumptions about the purposes of assessment, reinscribe problematic inequalities. This distortion happens in several ways. First, common assumptions about how to do assessment are guided by limited understanding of the nature of learning, as well as aligned with hierarchical and racialized beliefs about learners and learning. Deficit narratives and racial hierarchies undergird many widespread beliefs about assessment (Valencia, 2010). These tendencies make us uncurious about what is happening in schools, and about the patterns of inequity that we see again and again. This is ironic as standardized tests, because of their limited purview of learning, instead are more aptly understood as offering a representation of the policy compliance of a learning system—the classroom, the school, the district—and its relative health, including how it is serving different populations of students.

However, there are few diagnostics about learning processes that can help explain such outcomes that are happening within classrooms, schools, or districts. While some patterns of achievement, engagement and disengagement, or learning disparities can be surfaced by standardized tests across subject matters, how to understand these patterns is not made clearer through these outcomes because there is too much rich and contextual information missing. Thus, such assessments do not leave teachers or parents in a better place to support students,

nor to understand the right next steps for learners. And finally, what we are assessing is really about sorting (Oakes, 2005). This is evident in the ways we use the assessments: to sort students into categories of proficiency, to sort teachers into good and bad, and to sort schools into desirable and undesirable. As a society and as scholars of education, we should hold a bolder, more ambitious vision for assessment. We now turn to a description of some elements of that bolder vision.

How Should We Be Thinking About Assessment In Light of The RISE Principles?

These emerging big ideas around learning and development, captured in the evolving science of learning (SOLD) and summarized in what we call the R.I.S.E. principles, introduce radically new re-conceptualizations that stand in tension with the more siloed conceptions of learning and development that evolved out of the cognitive revolution. We focus on the evolutionary and biological foundations of human learning and development because these foundations urge attention to different aspects of learning and learning environments than traditional conceptions of learning.

Among these big ideas is the proposition that thinking and feelings, the emotional salience we attribute to experience, perceptions of the self, others and settings matter and operate in dialogical relations. Another big idea is that these dialogical relations unfold not in simple linear processes but in contexts of emergence. For example, Fisher, Frey, & Hattie (2016) documented trajectories of individual learners over time and found that the trajectories were not linear, and regressions may often be in service of development (Bever, 1982). The developmental research evolving around dynamic complex systems (Thelen & Smith, 1994) offers theoretical and methodological resources for studying how these dialogic processes of learning and development unfold within and across time.

One guiding idea in studying the complex systems of human learning in sociocultural and material contexts is the continuous mutual influence in real time among components as parts within a whole system. In the case of living organisms like human learners, this means organism–environment relations in an ecosystem (Rogoff, 2023). For example, we can conceive of learners' social interactions in or out of school as an ecosystem in which people act together, in concert, monitoring one another's actions, making next moves that take account of what others are doing. Another guiding notion from ecological system analyses

of the role of human interaction in learning is that participants in interaction use multiple sensory means to monitor one another's actions—not attending to speech alone through hearing but to visually and kinesthetically available information—and they draw on multiple semiotic resources in signaling meaning to one another in the everyday event timescales of microseconds, seconds, and minutes (Goodwin, 2017). Of course there are longer scale timeframes in play as well, as ecological relations exist between cultural practices inherited from our human ancestors are being redeployed in transformed ways to suit the learners' present circumstances (Lemke, 2000; Newell, 1994).

We also know that participation in routine cultural practices, social interactions with others, is central for learning. Further, the artifacts that human communities develop over time also matter for learning. For example, the field of epigenetics has demonstrated both that and how genes follow experience rather than the prior propositions, intellectually rooted in Eugenics, that genes determine human ability and possibility. We know further that among the essential targets of human learning and development from infancy on is what is called social cognition (Carlston et al., 2024), namely an evolving ability to read and respond interactively to the internal states of others. Studies in human infancy have documented both that infants pay more attention to other human beings than to objects and that infants and young children learn through observing and imitating human behaviors (Tomasello, Kruger & Ratner, 1993).

Research grounded in ecological systems theory highlights the significance of time and space in shaping human learning. Specifically, what and how people learn varies across different time scales—ranging from moment-to-moment (microgenetic) learning to developmental changes across the lifespan (Lemke, 2000). This perspective also considers how learning is channeled by the dynamic interplay among various settings and the cultural-historical contexts in which individuals live. For instance, Elder's (2018) longitudinal research on individuals who lived through the Great Depression demonstrates how learning and development across childhood, adolescence, adulthood, and older age were profoundly influenced by the specific resources and constraints framing that era.

Before we address the specific implications for assessments in schools, we should also address how the RISE principles require rethinking the ontology and phenomenology of the academic disciplines we teach within schools.

This rethinking will need to include how knowledge in the academic disciplines is operationalized in everyday contexts, including both the possibilities of connections between everyday knowledge and formal academic disciplines and the differences in such knowledge. Such connections are evident in the field of ethnomathematics (Rosa D'Ambrosio, Orey, Shirley, Alangui, Palhares, & Gavarrete, 2016) and in documented relations between indigenous epistemologies about the natural world and around what we think of as formal science (Medin & Bang, 2014). For example, Indigenous epistemologies concerning the natural world robustly conceptualize the interdependence among humans, animals, plants and other elements of the natural world. This framework stands in contrast to scientific epistemologies in biology that position humans at the top of a hierarchical ladder.

At the same time, there is evolving work in the biology field acknowledging humans as inter- dependently relational with the full breadth of the natural world (e.g., Seymour, 2016), and acknowledging the intertwining of genomes, biomes, microbiomes, and cultural meme pools (Leland et al., 2010). There is a growing understanding that humans are not only not separate from the natural world but are intricately connected and reliant on it for survival, with their actions affecting the environment and vice versa, creating a complex web of interdependencies. Humans rely on the natural world for resources like food, water, and air, and energy while simultaneously influencing the environment through activities like land use changes and energy transformation, and pollution. Ecological impact studies examine how human actions affect ecosystems and the biodiversity within them, considering the interconnectedness of all living organisms. Co-evolution names the idea that humans and the natural world have evolved together, with adaptations on both sides influencing each other over time. Biocultural studies combine biological and cultural factors to understand how humans interact with their environment, including their beliefs, practices, and social structures. Research around narrative sensemaking as an evolutionary disposition of humans (Bruner, 1990) is taken up to capture the diverse pathways through which storytelling (in everyday stories, in formal literature, in music lyrics, in digital media) is taken up across time and across cultural communities.

The point of focusing on the ontologies, phenomenologies, and epistemologies that inform academic reasoning across domains is to inform both the content of what we assess and the dimensions of learning for these domains. It is entirely possible

that an individual may demonstrate epistemological dispositions that are relevant to learning in a domain but not demonstrate adequate content knowledge.

The epistemological and phenomenological dimensions of learning are tools for building conceptual understandings. The ways in which assessments–formative or summative–can provide insights into such multiple dimensions of understanding is important. In some fields like mathematics, we have assessments that will reveal students' conceptual understandings, in part because in formal mathematics (from early to more advanced topics) the creation and manipulation of external representations of how one reasons is foundational to the field. Relative to science, standards such as the Next Generation Science Standards identify these multiple dimensions of scientific knowledge and reasoning, although there are not sufficient assessments available to address all the dimensions of knowledge and reasoning captured in the standards. In the field of literacy–reading, writing, vocabulary within and across disciplines–we do not have such protocols around external representations of processes of reasoning; and we thus tend in our assessments to capture outcomes, for example of comprehension or writing, but not cognitive or material processes of making sense of texts or writing processess. In contrast, when we observe typically on television programs or streaming media like YouTube (DeWitt et al., 2013)–sports programs, cooking programs, arts programs–when experts in the field observe others engaging in its practices, these media almost always de-construct the reasoning behind the individual's or the teams' decision making–making thinking visible in ways instrumental to learning.

What specifically do these big ideas tell us about assessing learning and development in the contexts of schools? What do they tell us about how we might engage in assessments of learning and development that unfold in contexts outside of schools, such as in family life or as people participate in activities within their broader community settings?

The Implications for Assessments

Because learning does not simply unfold inside the minds of individuals but more aptly occurs as described by the RISE principles, assessment systems need to provide windows into the multi-dimensionalities of learning and be ecologically valid. We focus first on systems of assessment, with the understanding that it is not merely what happens in classrooms that contribute to or constrain opportunities to learn. Such assessment systems should include windows into knowledge, the

learner, learning settings, and the organization of learning environments within and across settings (Barron, 2014). This framing builds on Gordon's (2007) notion of intellective competence, which contends that assessments should capture not only declarative knowledge, but also the "ability and disposition to use knowledge, technique, and values…to engage and solve both common and novel problems." Gordon's notion of values, and the importance of applying knowledge to not only familiar but also to novel problems, points to the non-linearity and multi-dimensionality of learning. It also points to the important higher-order ways of thinking that are crucial for adaptability and problem-solving.

Windows into knowledge include attending to and documenting the diversity of ontologies of knowledge, including conceptual, procedural, and/or epistemological forms of knowledge. Systems of assessment based on a more expansive view of learning can also offer insights into who the learner is without making restrictive assumptions about what is or not normative. In particular, these systems of assessments may include items that examine a learner's perceptions of themselves, their competencies, the learning settings they engage in, as well as perceptions related to their own coping, safety, and sense of belonging/connectedness to learning, to a discipline, and/or to a learning environment. Likewise, systems of assessment that are based on a multi-dimensional view of learning will also offer a window into understanding the learning settings in which the learner routinely operates that contribute to their learning. These settings can include family, community, classroom, and school settings, as well as specific policies and practices at district, state, and/or federal levels that can enable or constrain opportunities to learn.

Moreover, systems of assessment based on the RISE principles will be designed in ways that also examine how learning is organized in learning environments within and across settings and the opportunities that do or do not exist within them. Such assessments would include measures that capture opportunities for distributed engagement and exploration, not simply dominated by the teacher or whomever is assumed to have the greatest expertise relevant to the target(s) of learning. Similarly, these systems of assessment would facilitate opportunities for learners to create external representations of reasoning that make their thinking visible and/or allow learners to create and examine multiple modes of representations of reasoning as recommended by the Universal Design for Learning framework for learning materials (Rose, 2000). Finally, systems of assessment that hold the RISE principles

as foundational will recruit learners into accessing and utilizing their everyday repertoires relevant to learning tasks (e.g., language genres and registers; everyday applications of knowledge the learner may have experienced and explored outside of the current learning setting; epistemological orientations, particularly toward complexity and learners' experiences outside of the current setting of learning).

In arguing for ecologically valid assessment systems, we accept the premise that learning unfolds dynamically across settings and recruits multiple resources of the individual and the communities of practice in which the learner(s) engage. Thus, ecologically valid systems both provide windows into the elements of the system, as well as which ones are consequential for facilitating and/or foreclosing learning. For our purposes, crucial elements include the breadth of knowledge and dispositions that influential actors in the system, typically adults, deploy in supporting learning. In our formal education system, these key elements include the knowledge and dispositions, and indeed resources, available to teachers, instructional coaches, and specialized personnel such as social workers, counselors, and administrators at all system levels. Ecologically valid systems of assessment are needed in the United States, for example, as the distribution and quality of such knowledge, dispositions and resources are not equitably distributed. Among other issues, the U.S. is known for its problematic narratives about hierarchies of human communities, restrictive notions of learning as solely cognitive, and for a narrow scope of disciplines taught in school.

For example, LiPing Ma studies elementary school teachers of mathematics in the U.S. and China. Among a cohort of 5th grade teachers, she asked them to solve problems involving division of unlike fractions (Ma, 1999). Teachers from both countries could use the canonical procedure for solving such problems. However, when she asked them why they changed the operator from division to multiplication and inverted the numerator and denominator of the second fraction, not one of the U.S. teachers could explain why. Every Chinese teacher offered multiple conceptual mathematical explanations for why. This challenge involves more than the teacher's conceptual knowledge; equally crucial is what Lee Shulman (1986) described as pedagogical content knowledge—the teacher's understanding of what students need to know and do to engage in sophisticated disciplinary problem solving, the typical difficulties learners face, and the instructional strategies that can support their learning during the knowledge development process.

This difference between the cohort of U.S. and Chinese elementary math teachers is not explained by individual differences, but rather by the systems in place to support robust learning. When asked for her explanation of the differences, Ma explained that in China new teachers are not thrown into the classroom. Rather, new teachers work with master teachers in their school building who collaborate with them in planning, teaching and assessing. This parallels, for example, learning in medicine (Cooke et al., 2010). Graduates from medical school are not simply expected to make diagnoses on their own. Rather they work in long term internships to learn to apply what they learned in theory to practice with real and diverse human beings. We do not have such models of teacher learning either in schools of education or in school districts–with some rare exceptions (e.g., Bank Street: Nager & Shapiro, 2007).

Because we have argued for the breadth of what assessment systems need to provide windows into exploring, this means that those who administer and interpret findings from such assessment windows must have a breadth of knowledge to interpret findings from such assessment tools. These understandings include knowledge of child, adolescent and adult development, knowledge of the multiple dimensions of knowledge construction, and a deep disciplinary knowledge of what is being taught and assessed and how it can be fostered during instruction. These actors include teachers, instructional coaches, counselors, school and district administrators, including members of boards of education. All of these actors do not need the same depth of understanding in each area. For example, the school counselor or member of the board of education does not need deep conceptual understanding of the mathematics being taught by teachers and instructional coaches; but they do need to appreciate the fact that restrictive assessments of procedures and outcomes (such as reading comprehension assessments that only address comprehension outcomes but neglect the cognitive and social processes by which students go about making sense of texts) will not provide them with the kind of consequential knowledge on which to make ecologically valid instructional decisions.

While this agenda may sound overwhelming, there are exemplars of systems of assessment—including systems of assessment for preparing teachers and other actors in the educational system—that encompass the breadth needed to evidence an expansive, multi-dimensional view of learning. We offer the cases of OECD's Programme for International Student Assessment (PISA: Seitzer et al., 2021) and the Japanese Lesson Studies (Lewis et al., 2009).

Exemplar: Program for International Student Assessment (PISA).
We offer the PISA case as an example of how to design a program of assessment that does more than consider outcomes—identifying what works in a system and offering insights as to what needs to be changed. This is particularly important because across large-scale national U.S. data sets like NAEP and international comparisons from PISA and Progress in International Reading Literacy Study (PIRLS), we continue to see socioeconomic status and race/ethnicity associated with historically situated disparities in performance outcomes. We offer PISA as a contrast to the NAEP, which is the only national K–12 educational assessment in the U.S. NAEP assesses reading, mathematics, science, history, and civics in grades 4, 8, and 12 and reports levels of proficiency for knowledge outcomes in these content areas.

NAEP spans beyond student outcome reporting—it also issues surveys to teachers, administrators, and students. One aim of this broad reach is to document opportunities to learn (e.g., resource allocations, instructional practices), including surveying students about how they perceive each content area. Nonetheless, NAEP surveys are far more limited than those used in PISA because PISA also asks students about their sense of well-being and connections to school. PISA goes beyond cognitive outcomes to attend to social and affective well-being. OECD takes an ecological systems approach to data gathering, analysis, and understanding trends in social disparities around educational equity. In this way, PISA more clearly aligns with the expansive dimensions of learning and development discussed in this chapter.

PISA focuses on group trends over time nationally and, in the case of PISA cross-nationally, as a function of periodic administration to targeted population groups. In this way, PISA captures performance at varying grade and age levels and how those performances change over time. In addition, PISA examines the relationship of these performances to postsecondary outcomes, including participation in higher education and the workforce. PISA does not rely on a single assessment but draws from multiple assessments and surveys to make inferences about longitudinal patterns. These inferences, however, are not about the same populations or sets of students, rather, the large-scale dataset allows for size comparisons across data at different time points in the same participating nations. In this way, the assessments offer an opportunity to infer broad longitudinal trends.

Beyond reporting out about proficiency outcomes, OECD also generates a social disparity report. For example, 2018's PISA Social Disparities report examines how socioeconomic status affects learning outcomes across participating nations and the various factors for these differing outcomes (Organisation for Economic Co-operation and Development, 2018). The main PISA assessment program for 15-year-olds also includes indicators of students' sense of self-efficacy, sense of belonging in schools, effort and perseverance, career expectations, and measures of both concentrations of economic disadvantage and disciplinary climate in schools (Organisation for Economic Co-operation and Development, 2018). Analyses explore how equity in students' well-being has evolved as well as the extent to which disadvantaged students are socially and emotionally resilient.

The PISA 2018 report also includes a longitudinal examination of data from the Trends in International Mathematics and Science Study for data on fourth grade students as well as the Survey of Adult Skills, a product of the Organisation for Economic Co-operation and Development (OECD) Programme for the International Assessment of Adult Competencies (Organisation for Economic Co-operation and Development, 2018). This case illustrates what it takes to develop broad-scale national systems of teaching and assessment that provide the types of deep and wide scope of data that can be analyzed to better understand and explain variation in learning outcomes.

What we see in the PISA data is that the consequences of social background on educational success vary greatly across countries. Results from countries like Estonia, Hong Kong, and Vietnam also demonstrate within-country variability, wherein students who may be presumed to be at risk of failure instead succeed. Across OECD countries, more than one in ten disadvantaged students on average were among the top quarter of achievers in science (op. cit., p. 3). These data also suggest that the poorest students in one region might score higher than the wealthiest students in another country. Extending beyond the learning patterns made visible by the NAEP data, PISA's measures make clear it is not inevitable that disadvantaged students will perform worse than more advantaged students. There are positive contexts in which this result does not occur, invite the study of what Gawande (2007) aptly calls 'positive deviances'. The report concludes with a call for a broader understanding of learning, how learning environments affect learning, and for greater attention to the experiences of disadvantaged students in particular:

> *"Countries can also set ambitious goals for and monitor the progress of disadvantaged students, target additional resources towards disadvantaged students and schools, and reduce the concentration of disadvantaged students in particular schools. They can also develop teachers' capacity to identify students' needs and manage diverse classrooms, promote better communication between parents and teachers, and encourage parents to be more involved in their child's education. Teachers and schools can foster students' well-being and create a positive learning environment for all students by emphasizing the importance of persistence, investing effort and using appropriate learning strategies, and by encouraging students to support each other, such as through peer-mentoring programmes".* (Organisation for Economic Co-operation and Development, 2018, p. 15)

Lest we misrepresent PISA as a silver bullet, we must also acknowledge that PISA has been criticized as privileging developing countries and not adequately addressing issues of cultural relevance of content (c.f. Sjøberg, 2016; Teltemann & Klieme, 2017). Even with these critiques, OECD's efforts to address systemic features of educational systems that contribute to PISA outcomes are worth investigating as the data they gather spans far beyond NAEP's current scope.

Exemplar: Japanese Lesson Study.

Broad scale assessments like PISA can be helpful at a high-level to identify and offer insights into that which works in a system and that which needs to be changed. However, for individuals within a system, such as teachers, to learn to navigate that which is working well and that which needs to be changed, there also needs to be systemic support and tools for inquiry. To this end, we offer the example of Lesson Study in Japan—where teachers in school-based communities research their own practices and build this level of investigative and responsive practice into their daily workload and school day (Fernandez & Yoshida, 2004; Lewis et al., 2006).

Lesson study is a form of teacher education widely spread throughout Japan. It was introduced into the U.S. in the late 1990's by education scholars (Stigler & Hiebert, 2009) and was quickly taken up in the early 2000's by mathematics education scholars. While Lesson Study is often woven into pre-service teacher's methodology courses, lesson study has also increasingly been used by in-service

teachers who want to observe, discuss, and improve their pedagogical practices, classroom activities, students' learning experiences, and students' learning outcomes.

There are various forms of Lesson Study. At a top-level, the Lesson Study approach is concerned with how teachers "collaboratively plan, observe, and analyze actual classroom practice" (Lewis, Perry, Hurd & O'Connell, 2006, p 273). Specifically, the purpose of Lesson Study is to construct, through collaboration and observation, a practice-based theory that can be used to study and improve the teaching and learning occurring within a learning environment (Katakami, 2011). The process of Lesson Study can be engaged in and led by students, caregivers, teachers, administrators, and/or scholars, in various combinations. It does not require a top-down approach led by those with the most authority or power within a learning eco-system. Instead, lesson study is a collaborative process that can be engaged in by anyone invested in improving the skills, knowledge, and practices of teachers while also improving the knowledge base of teachers and the teaching profession (Fernandez & Yoshida, 2004).

The process of lesson study is not completely scripted, yet there is a general set of steps we will describe below (Fernandez & Yoshida, 2004). The first step is the collaborative planning of the study lesson. This is an opportunity for those involved to share their ideas about what the lesson should cover, how it should be designed, and what its learning objectives are. This planning draws upon the past teaching and learning experiences of the lesson designers, their understandings and observations of the current group of students who will engage with the lesson, their experience with the curriculum, and so forth. This collaborative process results in a lesson design that will be used to anchor the lesson study. While not explicitly identified in the Lesson Study literature, even this first step of lesson design deeply aligns with the multi-dimensional view of the RISE principles. This alignment can be seen in how Lesson Study takes seriously the social, cultural, emotional, cognitive, and contextual factors of a learning environment, as well as those of the students and teachers in that environment, to be ecologically valid and, ultimately, successful in achieving its learning goals.

The second step in Lesson Study is implementation (aka 'enactment') coupled with observation. This process involves one of the teachers actually teaching the co-developed lesson to their students, while the other members of the collaborative

design group observe the lesson's enactment. The public nature of this enactment requires the observers to also know the lesson well and to use it as a tool to guide their observation, note-taking, and subsequent reflection. In step 3 of the Lesson Study process, the design group reconvenes to reflect and discuss how the lesson unfolded. Each person involved, including the teacher(s) who taught the lesson, shares their observations and reactions to watching and/or engaging in the lesson. This review includes making suggestions about how the lesson could be improved vis a vis how it was implemented and experienced, how opportunities for learning were or were not presented and to whom, how it did or did not achieve the expected learning engagements and/or goals, and so on.

The fourth and fifth steps of Lesson Study involve revision and reteaching. While some design groups may decide to end their work at the third step and allow individual teachers to take it from there, design groups may choose to continue to learn together by building off of the reflections and suggestions to create (step 4) and teach (step 5) an updated version of the lesson design. The re-teaching of the new version of the lesson (step 5), mirrors that of step 2 wherein one group member, presumably a teacher, will again teach the lesson while their colleagues observe. Given the organization and nature of schools, it is unlikely that the same teacher will teach the same lesson to the same group of students a second or a third time. Instead, there is some variability with how the re-teaching occurs and who does it. This is an important aspect of Lesson Study, because the point is not to perfect a lesson nor to study the outcomes on a particular group of students to measure the success of a lesson, because the group values giving as many teachers as possible an opportunity to practice teaching the lesson and cultivating a broader base of experiences from which to learn and grow. These communitarian values are consistent with Lesson Study's purpose—to develop a practice-based theory relevant for studying and improving one another's teaching, as well as improving the learning and learning experiences of students.

The final and last step of Lesson Study, should the lesson be retaught, is to again share observations, reflections, and suggestions about the updated lesson version. As with all steps of the Lesson Study process, especially step 4, it is imperative that detailed notes are taken to document the discussion, the ideas generated, and the decisions, including their rationales. Such documentation offers a useful record for later reference should additional revisions and discussions accrue. It also is a necessary record should the teachers decide to report out or share insights about

their collaboration, their revisions, their enactments, etc. with others, whether at their school or in other professional settings concerned with improving teaching and learning.

These two exemplars—Japanese Lesson Study and PISA—suggest several key conditions that must be met for assessment systems to best support robust learning and teaching:

- Assessment developers need to understand human development, cultural communities, disciplinary knowledges, etc. to create effective assessments.
- Assessments need to include longitudinal and culturally-situated assessments of learners. Such assessments need to leverage multiple modalities, not simply texts.
- Assessments need to include social-emotional learning processes and outcomes; and finally,
- Assessments must examine opportunities for learning within learning environments and not only learners.

We have focused here on systems of assessment, yet there are also exemplars of specific assessment tools that individually address dimensions of learning we have identified. These tools are typically not widely distributed nor used. We cannot do justice to them within this chapter and instead invite those interested in these tools to look at those instances explored and discussed by Goldman and Lee (2024).

We will end this chapter with a discussion of some additional factors and tools that are worth considering in light of the current socio-technical context that influences education and many other societal sectors. In particular, the next section explores the potential that emerging technologies, such as generative AI, have for supporting the development of assessment tools that align with the RISE Principles of learning.

The Prospects of Generative-AI Augmented Assessment

In the emerging socio-technological universe, there is ample enthusiasm for the prospective roles of Generative AI in education's future. Generative AI developments based on large language models (LLMs) are proceeding at an unprecedentedly accelerative pace, affecting virtually every sector that produces learning media artifacts such as alphanumeric text, images, sounds, videos, as part of its information and knowledge production processes—which can now serve as input to large language and image models used to empower further generative AI advances. Researchers (Bick, Blanding, and Deming, 2024) and leading technologists and historians are comparing generative AI to the printing press and other epochal innovations like the World Wide Web.

Accordingly, entire industries are in dramatic transformational states with new companies further developing or exploiting generative AI technologies being funded with tens of billions of dollars. The future of human work itself is in question, which implicates some of the many purposes of education. Increasingly, the education sector is also attracting startup funding for generative AI applications and this trend will likely continue. While we cannot treat these opportunities and risks in any detail here, we can nonetheless point in the directions of inquiry which we anticipate will become promising for those enlisting Generative AI tools to augment our traditional socio-technical practices of K–12 learner assessments. We will begin with a brief preamble on how the RISE principles generally relate to the integration of Generative AI into K–12 assessments before describing three central reasons that Generative AI will be integral to the future of assessment.

As we've argued in this chapter, the RISE principles offer a way to view learning that honors its expansiveness and multiplicity and makes clear how social and cultural contexts are central to the learning process. Yet, how might these principles, and the research undergirding them, relate to the emerging uses of Generative AI for assessing learning? First, generative AI applications employed for assessment of learning congruent with the RISE Principles would need to take account of the expansive conditions of learning, as compared to today's more limited assessment paradigm of *assessing students inside school classrooms and class times* using standard assessments *that are either paper-based or computer-based* in their administration.

The understanding of learning being *Rooted* in the evolutionary, biological and neurological systems of our bodies and minds, and inseparable from our social and cultural activities, suggests that for emerging AI-augmented assessments to be effective, they would need to accompany a learner as they engage in and navigate a variety of socio-cultural activities no matter where they occur. Because learning is *Integrated* with all other aspects of development, including cognition, emotion, and the formation of one's multifaceted identity, to establish a wide-angle view of the whole child, emerging AI-augmented assessments will need to encompass these diverse fields of human functioning and not be restricted to the cognitive domain as they largely are today. This expansion would include such assessments being able to engage a learner's social and emotional states such as safety and belonging.

Due to learning being *Shaped* by everyday life cultural activities, both in and out of school, emerging AI-augmented assessments would be persistent throughout the daily activity rounds of the learner, not only in but outside of school contexts. Moreover, learning is *Shaped* by these activities across the lifespan, therefore assessments could not be one-offs that occur at only one given point for a very limited time, rather they would need to be longer-term and longitudinal to some degree (e.g., across a series of days, weeks, months, years, etc.). Lastly, since learning is *Experienced* in our bodies through coordination with social others and the natural and designed worlds, emerging AI-augmented assessments would more thoroughly engage with the physical bodies, gestures, and social and environmental interactions of learners as they coordinate their ongoing activities inside and out of school, including online virtual spaces. Biometric sensing data and multimedia records of learning interactions during activities in and out of school are likely sources of new assessment-related evidence for learning processes and outcomes, which will require satisfying all appropriate data privacy policies and safeguards.

Three central reasons provide further grounds to believe that Generative AI will be integral to the future of assessment congruent with the RISE principles of learning.[1] The first reason is that thoughtful uses of emerging and rapidly developing generative AI approaches *can yield improved assessment for use across traditional and new purposes.* These emerging approaches illustrate prospects for a dramatic

1 This section builds on text notes co-created by Eva Baker and Roy Pea for their work co-chairing the AI in Education Planning Committee of the National Academy of Education during 2024–2025.

change in the type and complexity of assessment tasks and modalities, because they transition the education research field from easily scorable multiple-choice and survey approaches toward tasks more closely approximating the situations, settings, and motivations involved in more complex learning environments. Examples include extended performance tasks with multiple dimensions, to include estimates and remedies for differences in prior knowledge (which is a persistent equity concern) and a range of cognitive demands, e.g., analysis, problem-solving (Anderson & Krathwohl, 2001; Mayer, 2009).

Recall that assessment practices often build upon Benjamin Bloom's cognitive taxonomy of educational goals and objectives (Bloom & Krathwohl, 1956), which proposed a pyramidal series of increasingly complex fundamental cognitive functions: remembering, understanding, applying, analyzing, evaluating, and creating knowledge. Further improvement likelihoods with AI are enhancing the assessment's relevance to the assessed individual's cultural and other experiential background details and preferences (Bailey & Duran, 2019; Bennett, 2023; Duran, 2020) and metacognitive processes (Fisher, Frey & Hattie, 2016; Yaden et al., in press) including components such as attention (Schwartz & Plass, 2020), motivation, and self-efficacy (Rueda, 2013; Rueda, O'Neil & Son, 2016) that can be readily monitored during the process of assessment and learning with learners' uses of technology-based systems. Congruent with the RISE principles emphasis on lifewide, lifelong and life-deep learning, AI is likely to enable greater empirical attention to affirmative development (Gordon and Bridglall, 2006), following individuals' intellectual growth over time within and across subject matter domains, inside and out of school, to ascertain patterns that can be supported or interrupted to produce outcomes of most value. At present, evidence of the developmental performance of individuals over time and contexts is rarely accessible. Indicators of development could very well be integrated across types and goals of various assessments, rather than limited to a particular type of task or format.

The second reason why Generative AI will be integral to the future of assessment is that *learning will be central to all emerging forms of assessment, when such assessments are re-situated as an integral aspect of learning,* rather than separated from it (Gordon Commission on the Future of Assessment, 2013). This orientation extends far beyond usual views on formative evaluation, which depend upon action taken from learning the results of tests or assessments. Instead, new affordances by large language models using artificial intelligence

(AI) can allow the learning and assessment activities *to be functionally blended from the learner's perspective,* while differentiating them in ongoing analyses. To take one example, simulations and games currently exhibit modestly blended elements (albeit highly structured ones: Schwartz & Plass, 2020; Shute et al., 2021). Learning activities with real-life problems based on domain knowledge will permit continuous and seamless feedback, adaptation, and learner adjustment in individual or collaborative settings, especially when accompanied by mobile platforms such as smartphones for the learners' activities which are contributing to their learning and to their being assessed.

It is very likely that education policy will continue to seek out or require summary documented evidence of learner growth or learning program and policy effects. AI will help enable the acquisition of such data to be sampled from individuals' ongoing learning in contrast to current, distinct, and more ceremonial assessments. Clear attention to learning will require tools for use by learning platform purveyors as well as smart support for teachers' own classroom assessments. AI can provide ways to integrate assessments from disparate sources that use common elements in their designs. Approaches to data collection, reporting, and validity and quality will also undergo substantial change. These modifications will require planning and systematic judgment to assure that AI and assessment together support the growth and success of all students.

Third, generative AI *can address barriers to equity found in current approaches to assessment and learning by addressing learner diversity in new substantive ways.* Equity should be a featured goal of AI and assessment, by supporting the development of insights into the full spectrum of the multitudinous ways in which individuals differ in their interests, knowledge, skills, dispositions, affect, motivations, and other learning-relevant characteristics such as bodily engagement. The goal is to capture patterns of performance, to supply needed background and more targeted prior knowledge or resources, and to incorporate appropriate information about learner preferences, experiences, and aspirations at the individual level for their more consequential educational support than is common today. Computationally mediated learning environments including interactive texts, symbols, graphics, audio, video, and animations will take adaptation far beyond current capabilities. Including contextual cultural elements in task structure along with desirable lexical cues can make learning and

assessment situations both comfortable and challenging for diverse learners. How such options should be developed and deployed will be a continuing scientific, policy, and values conversation and played out in the establishment of federal and state research and development priorities amidst a changing landscape of science and educational policies.

To be clear, AI already exists in assessments, particularly in writing (e.g., Ke & Ng, 2019). However, the present state of the art in AI and formal assessment and testing (e.g., OECD, ETS, Duolingo) attends to the use of AI mostly to improve task or item generation and scoring, yet for the most part uses existing measurement models that center on scalability. Importantly, changes in assessment scope and depth congruent with the RISE learning principles as sketched out above will necessitate the development of new approaches to common quality expectations for assessments. How will validity be ascertained for developmental performance once the metrics combine cognitive, affective, and domain-focused performances across learning contexts? How will reliability be reconceptualized when multiple items are not part of the assessment regimen? Will it be feasible to document assessment status for individual and policy use by using agents such as simulated students to reduce the response burden, time, cost, and delay associated with prototype testing and aggregated results for new assessments? It is certain that the infrastructure used to create learning and assessment designs and implementations will be upended. It will be vital to ensure that the clearest scientific knowledge and best practices in the learning sciences will be used to undergird these new AI-augmented learning and assessment systems.

However, such systems are not without their potential downsides. First, we know that AI systems are only as good as the data inputs they are trained on (Ferrara, 2023). Second, and relatedly, Generative AI systems have the potential to reproduce biases and deficit-oriented conceptions that exist in the broader society (Capraro et al., 2024). Hence, attending to the data inputs and the training of AI systems would need to be done carefully and thoughtfully, holding the RISE principles of learning and the ethical and equity-oriented cautions in mind.

Conclusion

We have argued in this chapter that to improve teaching and learning, and support robust learning, assessment systems in the US require a rethinking and revamping. We have offered one frame to support us in this redesign guided by the RISE principles of learning. The first principle argues that learning is *Rooted in the evolutionary, biological, and neurological systems of our bodies and minds, and inseparable from our social and cultural activities.* This suggests that assessment systems need to better account for the ways in which humans are, by definition, pack animals, who are motivated and learn in social contexts with others around them. The second principle is that *learning is integrated with all other aspects of development, including cognition, emotion, and the formation of identity—requiring for resolution a wide-angle view of the whole child.* This principle helps us hold at center how assessments and assessment systems need to account for issues of motivation, emotion, and connection, and not assume that one can separate the cognitive from other central developmental process. It also means that assessments must themselves be designed as experiences that motivate and engage. The third principle is that *Learning is shaped by everyday life cultural activities, both in and out of school and across the lifespan.* This reminds us that we must be vigilantly expansive in the ways that we design assessment systems, and that we should remember that schools are not the only, and perhaps even not the best) contexts within which learners gain important new knowledge and understandings. Ideally, we could draw on assessment systems to capture what schools tend to miss, and to provide points of leverage for supporting the integration of learning across learning settings as well as to include longitudinal dimensions to track learners over time. And finally, the fourth principle argues that learning is experienced in our bodies through coordination with social others and the natural and designed worlds. Thus, assessment systems, too, must draw on multimodal ways of learning and expression.

These pivots in how we design systems of assessment are critical if we are to move beyond the sorting function of assessment to build systems of assessment that are culturally inclusive and learner-centered, and which will provide important information about learner trajectories that can guide teaching in the future. In other words, we can most productively transform assessment when we build on the expansive understandings of learning that the sciences of learning and development make apparent. When we transform our understanding of learning,

and how to cultivate that learning, then the transformation of assessment systems should follow in alignment.

Importantly, at the heart of any assessment system must be a deep respect for the complexity of the learning process itself, and the intertwining of learning with a wide range of developmental processes and domains. We have also explored the possibility of AI augmented assessments which provide new opportunities to lean into some of these important properties of learning. It is a critical time in the field for these discussions, and perhaps more evident than ever that not only is it imperative that we develop more useful and more robust assessment systems, but that we utilize new tools and technologies based upon more robust understandings of learning to do so.

References

Abrahamson, D., & Lindgren, R. (2014). Embodiment and embodied design. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 358–376). NY: Cambridge University Press.

Alibali, M. W. (2025). Action, attention, and multimodal scaffolding. In Edwards, L.D., & Krause, C.M. (Eds.), *The Body in Mathematics: Theoretical and Methodological Lenses.* Boston: Brill.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Anderson, L. W., & Krathwohl, D. R., et al (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Bailey, A., & Duran, R. (2019). Language in practice: A mediator of valid interpretations of information generated by classroom assessments among linguistically and culturally diverse students. In S. Brookhart & J. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 46–62). New York: Routledge.

Baird, J. A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: fields apart? *Assessment in Education: Principles, Policy & Practice, 24*(3), 317–350. https://doi.org/10.1080/0969594X.2017.1319337

Bakhtin, M. M. (1981). *Four essays by M. M. Bakhtin.* In M. Holquist (Ed.). Austin: University of Texas Press.

Banks, J., Au, K., Ball, A.F., Bell, P., Gordon, E., Gutiérrez, K., Brice-Heath, S., Lee, C.D., Mahari, J., Nasir, N.S.,Valdés, G., & Zhou, M. (2007). *Learning in and out of school in diverse environments: Life-long, life-wide, life-deep.* Seattle, WA: LIFE Center.

Baroody, A. J., & Pellegrino, J. W. (2023). Assessment for learning. *Classroom-Based STEM Assessment*, 38.

Barron, B. (2006). Interest and self-sustained learning as catalysts of development: A learning ecology perspective. *Human development, 49*(4), 193–224.

Barron, B. (2014). Formative assessment for STEM learning ecosystems: Biographical approaches as a resource for research and practice. *National Research Council Committee on Out-of-School Time STEM.* Washington, DC: National Research Council.

Barron, B., Gómez, K., Pinkard, N., & Martin, C. K. (2013). *The digital youth network: Cultivating new media citizenship in urban communities.* Cambridge, MA: MIT Press.

Bennett, R. E. (2023). Toward a Theory of Socioculturally Responsive Assessment. *Educational Assessment, 28*(2), 83–104. https://doi.org/10.1080/10627197.2023.2202312

Bever, T. G. (1982). *Regression in the service of development.* In Bever, T.G., et al. (Eds.), Regressions in Child Development (pp. 153–188). NY: Routledge.

Bick, A., Blandin, A., & Deming, D.J. (2024). *The Rapid Adoption of Generative AI.* http://dx.doi.org/10.2139/ssrn.4965142

Bloom, B.S. & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: the classification of educational goals, by a committee of college and university examiners.* Handbook I: Cognitive Domain. New York, NY: Longmans, Green.

Bruner, J. (1990). *Acts of meaning.* Cambridge, MA: Harvard University Press.

Cantor, P., & Osher, D. (Eds.). (2021). *The science of learning and development: Enhancing the lives of all young people.* NY: Routledge.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., & Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS nexus, 3*(6).

Carlston, D. E., Hugenberg, K., & Johnson, K. L. (2024). Principles of social cognition: historical origins and current status. In D.E. Carlston, K. Hugenberg, & K.L. Johnson (Eds.), *The Oxford Handbook of Social Cognition, Second Edition* (pp. 3–19). Oxford: Oxford University Press.

Cole, M. (1996). *Cultural psychology: A once and future discipline.* Cambridge, MA: The Belknap Press of Harvard University Press.

Cooke, M., Irby, D.M., O'Brien, B.C. (2010). (Eds.), *Educating physicians: A call for reform of medical school and residency.* San Francisco: Jossey-Bass, 2010.

Darling-Hammond, L. (2010). *The flat world and education: How America's commitment to equity will determine our future.* NY: Teachers College Press.

Darling-Hammond, L. (2017). Developing and measuring higher order Skills: models for state performance assessment systems. *Council of Chief State School Officers.*

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2020). Implications for educational practice of the science of learning and development. *Applied developmental science, 24*(2), 97–140.

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., Bennett, R., Gordon, E., Haertel, E., Hakuta, K., Ho, A., Linn, R. L., Pearson, P. D., Popham, J., Resnick, L., Schoenfeld, A. H., Shavelson, R., Shepard, A., Shulman, L., & Steele, C. M. (2013). *Criteria for high-quality assessment.* Stanford Center for Opportunity Policy in Education, 2, 171–192. Stanford, CA: Stanford University.

Darling-Hammond, S. (2023). *Fostering belonging, transforming schools: The impact of restorative practices.* Palo Alto, CA: Learning Policy Institute.

DeWitt, D., Alias, N., Siraj, S., Yaakub, M. Y., Ayob, J., & Ishak, R. (2013). The potential of YouTube for teaching and learning in the performing arts. *Procedia-Social and Behavioral Sciences, 103,* 1118–1126.

Dryden-Peterson, S. (2016). Refugee education: The crossroads of globalization. *Educational researcher, 45*(9), 473–482.

Duran, R. (2020). Orienting Assessment toward Serving Students Reaction to Contributions of Edmund Gordon: We Need to Change Educational Practice and Assessment Together. *Educational Measurement: Issues and Practice.* 39. https://doi.org/10.1111/emip.12371

Elder, G. H. (2018). *Children of the great depression.* NY: Routledge.

Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci, 6*(1), 3.

Fisher, D., Frey, N., & Hattie, J. (2016). *Visible learning for literacy, grades K–12: Implementing the practices that work best to accelerate student learning.* Corwin Press.

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence,* 6, 1162454.

Flores, N. (2020). From academic language to language architecture: Challenging raciolinguistic ideologies in research and practice. *Theory into practice, 59*(1), 22–31.

Gawande, A. (2007). *Better: A surgeon's notes on performance.* New York: Metropolitan.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in cognitive sciences, 3*(11), 419–429.

Goldman, S. R., & Lee, C. D. (2024). Human learning and development: Theoretical perspectives to inform assessment systems. In Marion, Pellegrino, & Berman (Eds.), *Reimagining Balanced Assessment Systems.* DC: National Academy of Education.

Goodwin, C. (2017). *Co-operative action.* New York, NY: Cambridge University Press.

Gordon, E. W., & Bridglall, B. L. (2006). The affirmative development of academic ability: In pursuit of social justice. *Teachers College Record, 108*(14), 58–70.

Gordon, E. W., McGill, D., Iceman, S., Kelley M., Kalinich, Pellegrino, J. W., & Chatterji, M. (2014), Bringing formative classroom assessment to schools and making it count, *Quality Assurance in Education, 22*(4), 339–352.

Greene, J.L., Brock, C., Baker, W.D., & Harris, P. (2020). Positioning theory and discourse analysis: An explanatory theory and analytic lens. In Nasir, N. I. S., Lee, C. D., Pea, R., & McKinney de Royston, M. (2020). *Handbook of the cultural foundations of learning* (pp. 119–140). NY: Routledge.

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *The American Psychologist, 52*(10), 1115–1124.

Greeno, J. G. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53*(1), 5–26.

Gutiérrez, K. D., & Rogoff, B. (2003). Cultural ways of learning: Individual traits or repertoires of practice. *Educational Researcher, 32*(5), 19–25.

Heath, S.B, Bellino, M.J., & Winn, M. (2020). Adaptive learning across the life span. In Nasir, N. I. S., Lee, C. D., Pea, R., & McKinney de Royston, M. (Eds.), *Handbook of the cultural foundations of learning* (pp. 247–260). NY: Routledge.

Immordino-Yang, M. H. (2016). Emotion, sociality, and the brain's default mode network: Insights for educational practice and policy. P*olicy Insights from the Behavioral and Brain Sciences, 3*(2), 211–219.

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 6300–6308.

Kirst, M. W. (2024). *Standards-based education reforms: Looking back to look forward.* Palo Alto, CA: Learning Policy Institute. https://doi.org/10.54300/647.623

Kontra, C., Goldin-Meadow, S., & Beilock, S. L. (2012). *Embodied learning across the life span. Topics in cognitive science, 4*(4), 731–739.

Laland, K. N., Odling-Smee, J., & Myles, S. (2010). How culture shaped the human genome: bringing genetics and the human sciences together. *Nature Reviews Genetics, 11*(2), 137–148.

Lee, C. D. (2024). Implications of the Science of Learning and Development (SoLD) for Assessments in Education. In R.E. Bennett, L. Darling-Hammond, & Badrinarayan, A. (Eds.), *Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy.* Routledge

Lee, C. D., Meltzoff, A. N., & Kuhl, P. K. (2020). The braid of human learning and development: Neuro-physiological processes and participation in cultural practices. In Nasir, N., Lee, C., Pea, R., & McKinney de Royston, M. (Eds.), *Handbook of the cultural foundations of learning* (pp. 24–43). NY: Routledge.

Lemke, J. L. (2000). Across the scales of time: Artifacts, activities, and meanings in ecosocial systems. *Mind, culture, and activity, 7*(4), 273–290.

Lewis, C. C., Perry, R. R., & Hurd, J. (2009). Improving mathematics instruction through lesson study: A theoretical model and North American case. *Journal of Mathematics Teacher Education,* 12, 285–304.

Ma, L. (1999). *Knowing and teaching elementary mathematics.* Mahwah, NJ: Erlbaum.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge: Cambridge University Press.

McDermott, R., & Pea, R. (2020). Learning "how to mean": Embodiment in cultural practices. In Nasir, N., Lee, C., Pea, R., & McKinney de Royston, M. (Eds.), *Handbook of the cultural foundations of learning* (pp. 99–118). NY: Routledge.

McKinney de Royston, M., Lee, C., Nasir, N. I. S., & Pea, R. (2020). Rethinking schools, rethinking learning. *Phi Delta Kappan, 102*(3), 8–13.

McNeill, D. (2016). *Why we gesture: The surprising role of hand movements in communication.* NY: Cambridge University Press.

Medin, D. L., & Bang, M. (2014). *Who's asking? Native science, western science, and science education.* Cambridge, MA: MIT Press.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *The American Psychologist, 30*(10), 955–966.

Mislevy, R. (2018). *Sociocognitive foundations of educational measurement.* NY: Routledge.

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.). (2008). *Assessment, equity, and opportunity to learn.* NY: Cambridge University Press.

Murphy, R. F. (2019). Artificial intelligence applications to support K–12 teachers and teaching. *Rand Corporation, 10.*

Nager, N., & Shapiro, E. (2007). A progressive approach to the education of teachers: Some principles from Bank Street College of Education. *Occasional Paper Series*, *2007*(18), 1.

Nasir, N. S. (2000). "Points ain't everything": Emergent goals and average and percent understandings in the play of basketball among African American students. *Anthropology & Education Quarterly*, *31*(3), 283–305.

Nasir, N. I. S. (2024). A vision for the future of learning. *Educational Researcher*, 0013189X231222223.

Nasir, N. S., Lee, C. D., Pea, R., & McKinney de Royston, M. (2020). *Handbook of the cultural foundations of learning* (p. 480). NY: Routledge.

Nasir, N. S., McKinney de Royston, M., Barron, B., Bell, P., Pea, R., Stevens, R., & Goldman, S. (2020). Learning pathways: How learning is culturally organized. In Nasir, N., Lee, C., Pea, R., & McKinney de Royston, M. (Eds.), *Handbook of the cultural foundations of learning* (pp. 195–211). NY: Routledge.

National Academies of Sciences, Engineering and Medicine (2018). *How People Learn 2*, National Academies Press, Washington, DC.

Nathan, M. J. (2021). *Foundations of embodied learning: A paradigm for education*. NY: Routledge.

Newell, A. (1994). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Nichols, S. N., & Berliner, D. C. (2007). *Collateral damage: The effects of high stakes testing on America's schools*. Cambridge, MA: Harvard Education Press.

Niemi, H., & Nevgi, A. (2014). Research studies and active learning promoting professional competences in Finnish teacher education. *Teaching and Teacher Education*, *43*, 131–142.

Oakes, J. (2005). *Keeping track: How schools structure inequality.* New Haven, CT: Yale University Press.

OECD (2019), PISA 2018 *Assessment and Analytical Framework,* PISA, OECD Publishing, Paris, https://doi.org/10.1787/b25efab8-en

Organisation for Economic Co-operation and Development. (2018). PISA 2018: *Social disparities.* OECD Publishing.

Osher, D., Kidron, Y., Brackett, M., Dymnicki, A., Jones, S., & Weissberg, R. P. (2016). Advancing the science and practice of social and emotional learning: Looking back and moving forward. *Review of Research in Education*, *40*(1), 644–681.

Packer, M., & Cole, M. (2020). The institutional foundations of human evolution, ontogenesis, and learning. In N. Nasir, C. Lee, R. Pea, & M. McKinney de Royston (Eds.), *Handbook of the cultural foundations of learning* (pp. 3–23). NY: Routledge.

Pea, R. D. (1994). Seeing what we build together: Distributed multimedia learning environments for transformative communications. *The Journal of the Learning Sciences, 3*(3), 285–299. https://doi.org/10.1207/s15327809jls0303_4

Pinkard, N., Erete, S., Martin, C. K., & McKinney de Royston, M. (2017). Digital youth divas: Exploring narrative-driven curriculum to spark middle school girls' interest in computational activities. *Journal of the Learning Sciences*, *26*(3), 477–516.

Rogoff, B. (2023). Mutually constituting, fractal: individual and cultural aspects of holistic process. *Review of Research in Education*, 47(1), 60–83.

Rosa, J. D. (2016). Standardization, racialization, languagelessness: Raciolinguistic ideologies across communicative contexts. *Journal of Linguistic Anthropology*, *26*(2), 162–183.

Rosa, M., D'Ambrosio, U., Orey, D. C., Shirley, L., Alangui, W. V., Palhares, P., & Gavarrete, M. E. (2016). *Current and future perspectives of ethnomathematics as a program*. Springer Nature.

Rueda, R. (2013). 21st-century skills: Cultural, linguistic, and motivational perspectives. In D. Alverman & N. Unrau (Eds.), *Theoretical models and processes of reading*, 6th Ed. (pp. 1241–1267). Newark, NJ: International Reading Association.

Saracho, O.N. (Ed.), *Handbook of Research Methods in Early Childhood Education.* Charlotte, N.C.: Information Age.

Schwartz, R., & Plass, J.L (2020). Types of engagement in learning with games. In J.L. Plass, R.E. Mayer, & B.D. Holmes (Eds.), *Handbook of game-based learning* (pp. 53–80). Cambridge, MA: MIT Press.

Scriven, M. (1967). The methodology of evaluation. In R.W. Tyler, R.M. Gagne, M. Scriven (Eds.), *Perspectives of curriculum evaluation,* (pp. 39–83). Chicago, IL: Rand McNally

Seitzer, H., Niemann, D., & Martens, K. (2021). Placing PISA in perspective: the OECD's multi-centric view on education. *Globalisation, Societies and Education*, *19*(2), 198–212.

Seymour, V. (2016). The human–nature relationship and its impact on health: A critical review. *Frontiers in public health*, *4*, 260.

Shapiro, L., & Spaulding, S. (2024). *The Routledge Handbook of Embodied Cognition* (2nd ed.). Routledge. https://doi.org/10.4324/9781003322511

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational researcher, 15*(2), 4–14.

Shute, V., Rahimi, S., Smith, G., Ke, F., Almond, R. G., Dai, C. P., & Sun, C. (2021). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *Journal of Computer Assisted Learning*, 37(1), 127–141.

Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The Matrix of evidence for validity argumentation. *Frontiers in Education, 4*(43).

Steele, D. M., & Cohn-Vargas, B. (2013). *Identity safe classrooms, grades K-5: Places to belong and learn*. Corwin Press.

Stevens, R. (2020). Locating children's interest and concerns: An interaction focused approach. In Nasir, N., Lee, C., Pea, R., & McKinney de Royston, M. (Eds.), *Handbook of the cultural foundations of learning* (pp. 212–229). NY: Routledge.

Stigler, J. W., & Hiebert, J. (2009). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. Simon and Schuster.

Taylor, E. V. (2009). The purchasing practice of low-income students: The relationship to mathematical development. *The Journal of the Learning Sciences*, *18*(3), 370–415.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT press.

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and brain sciences*, 16(3), 495–511.

Turner, J. H. (2020). *On human nature: The biology and sociology of what made us human*. Routledge.

U. S. Department of Education, Office of Educational Technology (2023, May). *Artificial Intelligence and Future of Teaching and Learning: Insights and Recommendations*. Washington, DC.

Vakil, S. (2024). *Towards New Horizons of AI, Learning, and Equity in Education*. Spencer Foundation research report.

Valdés, G. (2004). *Learning and not learning English: Latino students in American schools*. NY: Teachers College Press.

Valencia, R. R. (2010). *Dismantling contemporary deficit thinking: Educational thought and practice*. NY: Routledge.

Volante, L., Klinger, D. A., & DeLuca, C. (2024). The rise and stall of standards-based reform. *Phi Delta Kappan*, *106*(2), 42–46.

Wang, T., & Cheng, E. C. (2022). Towards a tripartite research agenda: A scoping review of artificial intelligence in education research. In *Artificial Intelligence in Education: Emerging Technologies, Models and Applications* (pp. 3–24). Singapore: Springer.

Yaden, D. B., Rueda, R., Martinez, C., Mirzaei, A., Scott-Weich, B., Tardibuono, J., & Tsai, T. (in press). *A cross-linguistic perspective on studies using psycho-and micro-genetic design elements and methodology to assess young children's understandings of reading, writing and spelling.*

Implications of a Dynamic, Relational-Developmental-Systems Perspective for Research Design, Measurement, and Data Analysis in the Service of Understanding and Enhancing Youth Development and Learning

Richard M. Lerner and Pamela Cantor

Abstract

This chapter discusses dynamic, relational developmental systems models of human development, explains the specific value of these models for enhancing development, and presents the specific features of methodology actualizing this usefulness to address inequities associated with this variation for diverse groups and, in particular, the individuals within diverse groups. We provide specific examples of this variation across the kindergarten through high school span for diverse youth, and discuss the skills needed for them to thrive in both academic and other key settings of development within and beyond this span. These examples underscore the continuing inequitable relationships, experiences, and opportunities for diverse U.S. school children that, at this writing, exist within the American education ecology. These inequities can be addressed by identifying and promoting the specific instantiations of dynamic developmentally-nurturant relationships, that is, relationships that promote the health and psychosocial well-being of individuals across life periods and contextual settings.[1]

1 Richard M. Lerner's work on this chapter was supported in part by grants from the Templeton World Charity Foundation and Compassion International

> *"No existing form of life is truly solitary and no organism is completely independent of others at all times in its history. This dependence of every individual on others is the prerequisite to social behavior"* (1968, p. 68).
>
> Ethel Tobach, Curator Emerita, *American Museum of Natural History*
> T. C. Schneirla, Curator Emeritus, *American Museum of Natural History*

The purpose of this chapter is to discuss the implications of a dynamic, relational developmental systems (RDS) approach to human development for research methods aimed at describing, explaining, and optimizing the skills needed to flourish in and across age periods and settings. Specific examples of these implications will involve youth developing across the kindergarten through high school (K-12) span and the skills needed for thriving in both academic and other key settings for behavior and development within and beyond this span.

Given this purpose, we will illuminate the variation in individual-context relations that occur for diverse groups and, in particular, each of the individuals within diverse groups. The specific examples we will provide underscore the inequitable relationships, experiences, and opportunities for U.S. school children that, at this writing, continue to exist within the American education ecology. We will explain that specific instantiations of these relations can positively or negatively moderate the presence of *developmentally-nurturant relationships* (i.e., relationships that promote the health and psychosocial well-being of individuals across life periods and contextual settings). Examples of specific contextual setting are of course the family, school/classroom, and neighborhood/community. Relationships in these settings can then impact subsequent developmental trajectories across the like span. Some of these instantiations may create inequitable and socially unjust conditions impacts; such adverse circumstances are more likely to de experienced by diverse young people (Cantor et al., 2021; Cantor & Osher, 2021a, 2021b; Spencer, 2006; 2024; Spencer & Dowd, 2024; Spencer et al., 2015).

Accordingly, to address the purpose of this chapter we will first explain why dynamic and relational theoretical models of human life are of specific value for enhancing development–for instance, regarding achievement in school settings and in life beyond such settings. We will review the defining features of such models. This presentation will allow us to usefully discuss the specific features of

methodology–research design, measurement, and analysis–involved in actualizing the usefulness of dynamic, relational developmental systems models.

Why Dynamic, Developmentally-Nurturant Relationships are the Foundation of Flourishing Across Life

As illustrated by the epigram that opened this chapter, no form of life comes into being independent of a relationship with another life and, therefore, relationships between all individuals and their context have fundamental adaptive significance. Moreover, biological life is not only dependent on a supportive social relational context. It must also unfold in a life-sustaining physical ecology. As such, biological life is complex and dynamic; it is dependent on a system of mutually-beneficial social and physical ecological relationships that sustain both an individual *and* the other individuals and the physical ecology that are needed to sustain any person's life (e.g., Lerner et al., 2024-b; Lickliter, 2016; Li & Julian, 2012; Noble, 2015).

Accordingly, across human life and in all settings within which individuals behave and develop, such mutually-beneficial relationships among people must be present and must be maintained. For example, parents must act to create such relationships in their families and, as children develop, youth agency must be directed to contribute to these relationships. Similarly, students and teachers must have such relationships in classrooms, both as a matter of ethical and moral functioning and because such relationship can moderate the development of student cognitive, affective, and behavior skills (e.g., intentional self-regulation, perspective taking ability, theory of mind, and character virtues; Cantor et al., 2021; Lerner & Matthews, 2024). Opportunities should be provided for students' growing agency to contribute to maintaining and enhancing developmentally-nurturant relationships in school settings. Therefore, understanding the presence of positive or problematic behavior and development in a setting, for instance, a child in a classroom, must involve the integrated measurement of both the child and the setting and, in particular, must assess whether the measurement of student-teacher relationships in the setting provides evidence of mutually-beneficial, developmentally nurturant student-teacher relationships.

However, the complexity of the measurement of such dynamic, developmental relationships is particularly challenging for human beings. For instance, infants and their caregivers must be attuned to, and create a goodness of fit with each

other, if the extended periods of human infancy and childhood that are emblematic of human development, are to be optimally traversed and, as well, if these relationships are to produce health and well-being for all members of an individual's proximal (e.g., home, classroom) and more distal (neighborhood, community, national) ecological systems (e.g., Bronfenbrenner, 1979, 2005; Bronfenbrenner & Morris, 2006; Chess & Thomas, 1999; Thomas et al., 1963). As a consequence, measures must be longitudinal to assess the ongoing dynamics between children and parents.

In turn, if teachers are to avoid burnout and maintain the energy and motivation to monitor and maintain developmentally-nurturant relationships with all of the students in their classrooms, then across the K–12 span students must increasingly contribute to such dynamic relationships. Again, then, assessment of the existence of such relationships must involve longitudinal measurement and, as well, a determination that valid age-appropriate and contextually-sensitive measures are used across the 13 years of development involved in the K–12 span.

Simply, within and across all developmental periods and all contexts it is vital to measure and analyze information about whether mutually-supportive individual-context relations are present. If there are no developmentally-nurturing relationships among humans, then thriving is obviated for all people involved in these coactions and even life itself may be threatened (Gould, 1977; Johanson & Edey, 1981).

Of course, no two human beings have exactly the same history of social relationships and of experiences with their physical ecology across time and place (Elder et al., 2015; Hirsch, 2004; Moore, 2002, 2015). Nevertheless, amid such variation, it is still possible to specify the characteristics of human relationships and experiences that place all individuals on positive and healthy pathways across the life span. We briefly review such dynamic relational concepts next. Each of these concepts has implications for assessing the coactions among youth and the key contexts of their lives and, therefore, for illuminating the coactions among students, teachers, and their classrooms and schools.

Key Concepts Within Dynamic, Relational Developmental Systems-Based Models

The dynamic, relational development system within which all human life develops has been shaped by the evolutionary biology of the species (e.g., Gottlieb, 1992; Jablonka & Lamb, 2005; Johanson & Edey, 1981; Lickliter, 2016; Moore, 2015; Noble, 2015). The embeddedness of humans within this evolutionary context has necessitated that mutually-influential relations between an individual and the context–which we will represent as individual ⇔ context coactions–also involve mutually-nurturant relations for humans to both survive, per se, and to thrive within and across the life span. Three concepts describe the evolutionary distinctiveness within which developmentally-nurturant relations have evolved and that enable healthy and positive human life to exist.

Neoteny, paedomorphy, and the cultural dimension of evolution.

Humans are the most *neotenous*–that is slowly maturing–of all animal species (Gould, 1977). As compared to protohominid and hominid ancestors and to other animal species in contemporary existence with *homo sapiens*, humans take a decade or more to reach skeletal-, organ system-, and particularly reproductive-maturity. Second, because of their neotenous growth, humans are also the most *paedomorphic* of all ancestral species; that is, humans maintain child-like physical and functional characteristics longer in their life spans (their *ontogeny*) than any of their ancestral species or, again, other contemporary species (Gould, 1977). For example, new neuronal growth accompanies the adolescent period of development, and new brain myelinization is in evidence in the seventh and eighth decades of life (Baltes et al., 2006; Dorn & Beltz, 2022; Paus, 2009). What is most salient about these two distinctly human developmental features is that the coaction of neoteny and paedomorphy combine to make humans more dependent on positive nurturing from developmentally-attentive caregivers for longer in their life spans than any other animal across either evolutionary history or contemporary circumstances.

In addition, a third facet of human beings pertains not only to the singularity of their evolutionary history among all past and present organisms but, as well, to the complexity of human development itself. The evolutionary history of human life involves the integration of four dimensions of processes. These dimensions are genetics, epigenetics, human behavior, and culture. In particular, it is the fourth of these dynamically-interrelated (i.e., mutually-influential and mutually-regulated)

dimensions of human evolution that creates the uniqueness of human life (e.g., Raeff, 2016; Rogoff, 2003).

Together, the nature and course of human development is unique among all species because of this suite of features of human life—neoteny, paedomorphy, and the four dimensions of human evolution. It is especially critical that the fourth, cultural dimension of human evolution be appreciated when attempting to specify what makes humans human. The cultural dimension is qualitatively and quantitatively different from what exists for other species. Humans singularly possess complex cultural creations, such as language and other symbolic systems and, as well, educational, economic, political, and religious traditions and institutions.

Such features combine to define human life and its purpose and meaning differently from our evolutionary ancestors and from all other living creatures. In fact, these features help to reveal what is individual about each human being and to explain why and how human variation is the norm and not the exception in human development. For youth who are assessed for their academic (e.g., science, technology, engineering, and math) skills or life skills (e.g., socioemotional learning or character attributes), the features of human evolution we have noted require integrated longitudinal measurement of both individual and context and the analysis of such relational data. In addition, measurement of the characteristics of individuality of each student must be measured as well and, compared not only to other students or to school, district, regional, or national norms but, and as important, to the individual students themselves. Such analysis of the person in relation to themselves is a data analysis approach that is termed ipsative, person-specific, or idiographic. We will discuss such analytic methods in great detail later in this chapter, but here it is important to emphasize that, without including such measurement in the assessment of student ⇔ context relations, educators, parents, or students themselves cannot know all facets of the person that drive their behavior, capacities, well-being, or possible futures (Lerner, 2021; Rose, 2016). In addition, as we now discuss some of the other key concepts associated with dynamic, relational developmental systems ideas about humans, we will highlight that these concepts also have important implications for measuring individual ⇔ context relations (see too Lerner, 2018; Moore, 2015; Overton, 2015 for additional details).

Embodiment

Overton (2015) explained that human development is embodied by 1. the physiological and morphological features of humans (e.g., neurobiology, genetics/epigenetics, and hormones), by 2. the coactions of psychological processes (e.g., involving cognitions, affect, and behaviors) with this first instance of embodiment, and by 3. the coaction of social and cultural processes with these first two instances of embodiment. Thus, Overton (2013) notes that:

> *"Embodiment includes not merely the physical structures of the body but the body as a form of lived experience, actively engaged with the world of socio-cultural and physical objects. The body as form references the biological point-of-view, the body as lived experience references the psychological subject standpoint, and the body actively engaged with the world represents the socio-cultural point-of-view"* (p. 103).

As an example of this embodiment, Immordino-Yang and colleagues (e.g., Immordino-Yang et al., 2019; Immordino-Yang & Yang, 2017; Immordino-Yang et al., 2023) point to evidence indicating that the individuals' cognitive, affective, and behavioral processes involved in meaning-making of their world are a product *and* a producer of meaningful changes in brain activity that result from the individual coacting with features of their specific contexts. Thus, developmental and educational sciences require theory-predicated and empirically systematic interrogation and, ultimately, the integration of the three domains of embodiment.

As well, understanding the current status of a student in a classroom and, arguably more importantly, using such information to help predict the student's range of developmental possibilities and life attainments will require valid measurement of all domains of embodiment, measurement that is longitudinal, and data analyses that validly depict the range of possibilities for every specific student. Recognition of the importance of such methods of data collection, measurement, and analysis is underscored by yet another concept from dynamic, relational developmental systems models.

Holism

Human life and development must be understood as involving multiple domains of development that coact in the holistically integrated dynamic developmental system. The concept of holism means that the "parts of the whole" (the embodied domains) of an individual:

> *"do not combine through an additive process. Instead, the combination may be better understood as a multiplicative process: When the parts combine, they produce, in combination, attributes of a novel whole that do not exist in the parts in isolation. What makes living systems unique is that they change systematically, through mutually-influential individual ⇔ context relations, into new, increasingly adaptive and complex forms (see Gould & Vrba 1982; Jablonka & Lamb 2005). The whole, through its self-organization, has unique systemic features that are not attributes of any part. Thus, the whole is not just quantitatively greater than the sum of its parts; it is qualitatively different from the sum of its parts"* (Cantor et al., 2021, p. 23).

For instance, the key methodological point associated with holism is that measurement of a student without also measuring the student's context is *meaningless*. A purported measure of a student's current capacity or future potential that results in the assignment of a score attached to an individual is useless unless that score is empirically understood (through integrated measurement of the student and context that has occurred across time—that is longitudinal–and that includes valid assessments of all domains of embodiment) to reflect holistic assessment. A score attached to a student that is not understood holistically as a score of the students and context at one specific time is both completely inadequate and potentially damaging to the well-being of a student. Without integrated understanding of the dynamics between students and all the contextual influences impacts on them, students who have lived or now live in challenging contextual circumstance, for example, because of marginalization associated with racism, poverty, and minoritization, cannot be accurately or adequately assessed (Cantor et al., 2021; Card, 2017).

By ignoring the developmental contextualization of the academic scores of any student, educators will (continue to) unfairly penalize marginalized students and, in turn, privilege in unjust ways students who have lived free of unjust marginalization and minoritization (Spencer, 2006, 2024; Spencer et al., 2015). Accordingly,

educational and development science must use methods of assessment and evaluation that know each student in the wholeness of their individuality (Cantor et al., 2021; Lerner, 2021; Lerner & Greenberg, 2025; Lerner & Matthews, 2024; Rose, 2016). We return then to the concept of specificity.

Specificity of Individual ⇔ Context Relations

As we have noted, human life and development are characterized by the specificity of an individual's course of individual ⇔ context relations across time and place (e.g., Bornstein, 2019; Elder et al., 2015; Molenaar, 2004; Rose, 2016). It is unlikely that any two people will have the same history across their lives in these relations (Bornstein, 2019; Hirsch, 2004; Rose, 2016). Said another way, the history of individual relations or relationships that occur across a human life are always specific to a person (Bornstein, 2019; Rose, 2016).

The Specificity Principle presented by Bornstein (2017, 2019, 2024; see too Lerner & Bornstein, 2021) depicts the dimensions of biological, psychological, behavioral, relationship, and contextual individuality that must be understood to provide a complete (embodied and holistic) account of human development and, as well, to maximize the chances that the application of human developmental theory-based research equitably promotes positive outcomes for every individual. Indeed, without recognition of specificity in models and measurement of human development (e.g., Molenaar, 2004; Rose, 2016), and in evidence-based policies and programs derived from such developmental scholarship, neither authentic equity nor social justice can be interpreted as useful contributions from developmental or educational science (Cantor & Osher, 2021).

Of course, the specificity of individual ⇔ context relations does not negate the existence of facets of developmental processes that can be generalized across groups (i.e., differential group processes) and, as well, facets of developmental processes that are nomothetic (that are shared by all humans) (Allport, 1937, 1962; Emmerich, 1968). Kluckhohn and Murray (1953) made this point by explaining that it is simultaneously true that each human has attributes that are present in all other humans, that each human has attributes that are present in only some other humans, and that each human has attributes that are shared with no other humans (these attributes are labeled as *idiographic* or *person-specific*).

Implications of Dynamic, RDS-Based Concepts for Research Design, Measurement, and Analysis

In the enactment of good science, the methods selected for use in an empirical investigation should align with the questions being asked by the researcher. Indeed, methodologists studying human development have asserted that theory is the primary tool of developmental methodology (e.g., Collins. 2006). In other words, theory-predicated questions should guide the choice of methods (measures, designs, data analyses) used in developmental research. As noted in prior sections of this chapter, concepts of development associated with dynamic, RDS-based models of character development have specific implications for the questions asked in developmental and educational science research and, in turn, for the methods used to address the questions.

Neoteny and paedomorphy

Because of neoteny and paedomorphy, the assessment of individual ⇔ context relations must extend over large portions of the human life span to capture the changes that are involved in a person moving from birth to maturity. Accordingly, point-in-time measurement (e.g., as is studied in cross-sectional research designs) may be useful to address some empirical questions (e.g., the average scores for reading among students in Grades 2, 4, and 6 in a specific school or school district). However, only repeated measurement–longitudinal–designs can provide information about an individual's development (whether there are changes in individual students reading scores across Grade 2, 4, and 6).

Thus, cross-sectional designs (Baltes et al., 1977) only provide data about interindividual (between-person) differences within specific times of measurement, and these differences may or may not be due to developmental change or to treatment effects (e.g., Shonkoff et al., 2017). For instance, interindividual differences may reflect birth cohort differences and/or inadequately matched age groups (Baltes et al., 1977; Lerner, 2018). No data about intraindividual (within-person, ipsative, or idiographic) change can be derived from the point-in-time measurements gathered in cross-sectional studies. In addition, at least since the classic papers by Schaie (1965) and Schaie and Strother (1968), the ensuing 50+ years of developmental science research has repeatedly demonstrated that developmental trajectories derived from cross-sectional research do not align with developmental trajectories derived from longitudinal research. Simply, then, cross-sectional designs cannot provide

data illuminating the course of *development* across the life span (e.g., a growth mindset) among, or of enduring outcomes of, K–12 educational programs (e.g., the creation of a life-long learner; e.g., Baltes et al., 2006).

Embodiment

Because of embodiment, longitudinal measurement must involve measurement of both individuals and their contexts and, therefore, questions about the bases of a student's academic skills, likely pathways of educational and life attainments, or socioemotional or character attributes must interrogate the physical and physiological attributes of the student, the psychological and social characteristics of the student, and the social and cultural context of the student.

For instance, much of the data collected about positive attributes of human development (e.g., grit, character strengths, growth mindset, hopeful expectations about the future) solely involve surveys, a methodology that has well-known measurement problems (e.g., response bias, social desirability-based responses) in regard to establishing validity and reliability (Card, 2017; Clifton, 2020; Rioux & Little 2020). In addition, few survey measures of character have established measurement invariance across age, gender, ethnicity and, in particular, in regard to international generalizability, across nations and cultures (Putnick & Bornstein, 2016; Maasen et al., 2023; Tirrell et al., 2019). Furthermore, as explained by Nucci (2017, 2019, 2024), within dynamic models of human development meaningful change is a multifaceted, integrated, and dynamic developmental phenomenon and, as such, a narrow approach to measurement could not validly capture this complexity.

Accordingly, measurement models should include programmatic measurement of all three domains of variables comprising the embodied nature of all development processes, On the one hand, there has not been systematic assessment across the life span of the coaction between morphological and physiological processes with the specific manifestations of holistic thriving within person-specific processes. For example, are there specific times in K–12 education programs demonstrating effectiveness in promoting academic achievement, SEL competencies, and active and positively engaged citizenship? Are there specific tipping points for establishing life styles that, across decades, foster both physical and mental health and active stewardship of the planet that sustains human life and the well-being of future generations?

On the other hand, although there is considerable evidence about the link between sociocultural context and diverse facets of human development (e.g., Lerner et al., 2021a, Lerner et al., 2021b; Raeff, 2016), these links have rarely been extended to include coactions with the morphological and physiological dimension of embodied development processes. The result of these omissions is an impoverished understanding of the whole of a human being.

Holism

Because of holism, measures associated with the three levels of embodiment must be *integratively* assessed and analyzed for their statistical interactions and the possibility that measures of one or more of the attributes that are measured mediate changes in other scores. Much of the research literature on the development of thriving or well-being reports work about one or at most a few attributes. Indeed, single-topic scales (e.g., grit, growth mindset, curiosity) dominate research. In addition, many programs that undertake evaluations of their work focus on one or only a few attributes as outcomes of their interventions (e.g., see Lerner et al., 2021a and Lerner et al., 2021b for discussions).

As implied in our discussion of the methodological implications of embodiment, the absence of a holistic approach to all levels of organization constituting human life is complicated when only a narrow assessment of constructs representing each level is included in a measurement model. One basic tenet of statistical analysis is that when a main effect of a variable is embedded in a higher-order statistical interaction, interpretation of main effects is eschewed in deference to interpretations of the interaction effect. The joint ontological implication of embodiment and holism is that each human life is a result of a complex coaction among many variables within and across levels of integration, including, as we have emphasized, individual ⇔ context coactions (e.g., see Heckman et al., 2024). To understand the embodied and holistic role of thriving in the development of this complex system is a necessary, but admittedly daunting, challenge for developmental or educational science.

Of course, issues of feasibility, time, and financial resources limit the measurement model of any single study. Therefore, in order to design and implement research that can holistically measure individuals across time and place, a combined implication of embodiment and holism is that researchers must pursue a research

topic programmatically, perhaps ones that involve cross-laboratory efforts and are part of a collaborative national and international network.

At this writing, there are signs that support for such a program of scholarship in the study of the development of holistic thriving is being considered as a target of funding by some private foundations. The Templeton World Charity Foundation's Global Innovations in Character Development is a case in point (Dill & Simpson, 2024). The significance of this kind of integrative research will foster more complete definitions of health—and more important research questions such as what is a healthy person, how do researchers or educators define such health, or how do researchers or educators know that learning has been optimized in a given context and that a student is performing at the top of their developmental range?

Such integrative and programmatic funding will be more likely to appear prudent among funders if research about the embodied and holistic developmental system can be coupled with compelling evidence that, consistent with RDS-based ideas about the specificity of human development, the study of the specific complex and dynamic developmental course of individuals can be studied in the wholeness of their individuality (Lerner et al., 2022, 2024a; Lerner & Matthews, 2024), including the supports for the health and learning present on their contexts across life.

Developmentally nurturant relationships

Furthermore, the measurement model in such assessment must include indices of the presence or absence of developmentally nurturant relationships in the student's life and, more generally, of the impacts of contextual influences on every measurement made of the student. An open and dynamic system can be changed for the better or for the worse. The dynamic developmental system of a person manifesting holistic thriving is the same sort of dynamic system affecting a person manifesting problematic behavior across the life span. What, then, accounts for these interindividual differences?

We believe that the critical difference between a person who becomes an example of positive development or grows up to be a "menace to society" was famously stated by Theodore Roosevelt: "To educate a person in mind but not in morals is to educate a menace to society." The focus of education implied by Roosevelt is an instance of education targeted at a specific component of holistic human development (Nucci, 2024): Character education. This educational focus is a

topic that developmental scientists have studied quite extensively (Matthews & Lerner, 2024a, 2024b). These studies include a focus on formal (in-school) and informal (e.g., family socialization practices [Bornstein, 2024] or out-of-school time educational programs [Berkowitz & Baer, 2024]) a person experiences across the life span. Whereas this point may seem obvious, it nevertheless has methodological implications that will enhance understanding of the developmental bases of holistic thriving.

It is likely that, except for a few resolute genetic reductionists, that is, those writers who believe that the genes inherited at birth are a blueprint for all later development (e.g., Plomin, 2018), most developmental scientists would agree about the significance for shaping moral development and character strengths that exists within the formal and informal educational experiences that are part of a person's socialization. However, there have been far too few investigations of the processes that enable precise delineation of the specific formal and information educational experiences, and of the coactions among them, that are needed to promote specific features of character development among specific youth. Interrogating the conditions across life that eventuate in specific features of an individual's character attributes results, then, in framing empirical questions through the use of the Bornstein Specificity Principle (2017, 2019, 2024).

Specific instances of student ⇔ context relations

As we have already emphasized, because all scores attached to a student's academic record are not indices of the individual student, per se, but of specific instances of student ⇔ context relations, such dynamism must be measured and analyzed (e.g., see Deboeck et al., 2023). In addition, the insistence on such specificity does not deny that a comprehensive and integrative approach to measurement will reveal that, in the embodied measures of a student, some portion of the variation will reflect various facets of life present in all people, some portion of the variation will involve facets of life present in only some groups of people, and some portion of the variation will involve facets of life present only in one the specific person.

Bornstein's (2019) Specificity Principle involves programmatic developmental research that addresses a complex, multi-part question aimed at generating the evidence needed for precise knowledge of the bases of holistic positive

development. An example of the set of questions that may derive by framing research through use of this principle involves asking:

- What specific attributes of holistic thriving;
- For a person of what specific demographic or status characteristics (e.g., age, gender, race, ethnicity, religion, SES, etc.);
- Experiencing what specific formal or informal socialization (or education) experiences;
- In what specific setting(s) (e.g., family, school classroom, out-of-school-time setting, professional training, religious training);
- Of what intensity, duration, and engagement level;
- In what specific community, society, and culture;
- At what specific time in the life span;
- At what specific time in history;
- Will result in what manifestations of holistic thriving?

Research about holistic thriving must build into it change-sensitive measures suitable for use in addressing such a set of questions about the specifics of individual ⇔ context coactions that occur for specific youth, at specific times in life, and within and across specific settings. Developmental or educational research or program evaluations guided by such questions will enable holistic understanding of how each person can have a life span marked by mutually beneficial coactions between the self and a civil society.

In sum, although we believe that all researchers in developmental or educational science would likely acknowledge that all three instances of variance–variance associated with all humans (nomothetic variance), variance associated with only some people (group differential variance), and variance specific to the individual (idiographic variance)–inevitably exist in any human data set, not all of these researchers would insist that all three sources of variation are meaningful (e.g., Hamaker et al., 2018; Molenaar, 2004; Molenaar & Nesselroade, 2015; Rose, 2016). Whereas many researchers might contend that individually-specific fluctuations around population or sample means constitute random variation of error variance (e.g., Rose et al., 2013), the concepts that we have discussed as associated with dynamic, RDS-based models may lead a researcher to believe that such variance is

meaningful and requires the use of analytic statistical methods, such as dynamic structural equation modeling (DSEM; McNeish & Hamaker, 2020) or state space grids (e.g., Hollenstein, 2007) to test such beliefs. Because such interpretations of person-specific data are still relatively new and controversial at this writing (see Deboeck et al., 2023; Hamaker et al., 2018), it is useful to discuss in more depth this facet of the methodological implications of RDS-based concepts in more depth.

The Importance for Developmental and Educational Science of Measuring and Analyzing Person-Specific Data

Over the past two decades, substantial methodological innovations have been made in the person-specific study of human development. These innovations have not ignored or attempted to explain away through ontological reductionism the complexity of person-specific development or to treat nomothetic or differential variation as a "good enough" approximation of idiographic variation (Lerner, 2018, 2021). Quite to the contrary, the methodologists leading the way in person-specific measurement and data analysis have fully embraced such complexity (e.g., Deboeck et al., 2023; Hamaker et al., 2018; Mascolo & Bidell, 2020; Molenaar, 2004; Molenaar et al., 2014; Nesselroade & Molenaar, 2010; Ram & Gadzke-Kopp, 2023; Ram et al., 2005, 2014; Rose, 2016).

A useful reference point for the emergence of an emphasis on the importance of focus on the individual *qua* individual in pursuing a holistic and dynamically integrated understanding human development is a 2004 article authored by Peter C. M. Molenaar, "A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever." Molenaar (2004, p. 202) noted that "Attention in psychological research is almost exclusively restricted to variation between individuals (interindividual variation), to the neglect of time-dependent variation within a single participant's time series (intraindividual variation)."

Indeed, writing more than a decade later, Hamaker et al. (2018, p. 820) observed that, when their *intensive* (i.e., densely-sampled observations across relative short periods of time) longitudinal research focused on within-person change begun (at about the time that Molenaar published his Manifesto), "gathering intensive longitudinal data was not only cumbersome, but it was also considered unnecessary by many, because short-term, within-person fluctuations were

assumed to reflect mere noise… [and not] the meaningfulness of short-term, within-person fluctuations." Similarly, McNeish and Hamaker (2020) explained that:

> *"Developmental process data… typically feature a few measurement occasions that are widely spaced over the observations window (often months or years apart). The primary focus is questions about the means of the outcome variable over the course of the observation window… how much do the means change from the first to the last measurement occasion?… These models tend to take a nomothetic approach in describing the overall mean across people…Moreover, when covariates are added to the model, the predominant focus is variables that affect the shape of the growth curve… Plainly stated, covariates for developmental process commonly explain between-person variability (factors that lead to differently shaped growth curves) but less commonly explain within-person variability (deviation around the growth curve)"* (pp. 611–612).

In essence, then, to understand holistically and dynamically integrative developmental processes of development, the study of an individual's repeated scores across time (i.e., the study of within-person, intraindividual change) must be combined with the study of variation in averages (nomothetic or differential changes). Moreover, and as we emphasized throughout this chapter, within dynamic, RDS-based models, holistic and dynamically integrated understanding of individual pathways of development must focus on integrated individual ⇔ context coactions, and not on either individual or context alone. Indeed, the necessity of understanding coactions with the specific setting or context of the individual in order to understand the course of the person's development is formally specified through Equations 1 through 5 in a provocative and convincing chapter by Heckman et al. (2024).

In short, there is rich and important support of the Molenaar (2004) Manifesto call for the idiographic study of the process of development and learning in general and, more particularly, for the importance of adding to the field a focus on the ways in which specific, dynamic individual ⇔ context relations are a foundational part of *all* developmental and educational processes (e.g., Cantor et al., 2019; Lerner et al., 224; Lerner & Matthews, 2024; Matthews & Lerner, 2024a, 2024b; Moore, 2015).

Certainly, this focus on the necessity of understanding dynamic individual ⇔ context specificity has been elevated by the influential scholarship associated with the Bornstein Specificity Principle (e.g., 2017, 2019; 2024). At this writing, this concept has attained the status of a foundational principle of developmental science. It should also be seen as a key idea within metatheories and theories used in educational science.

In the medical sciences, thoroughly understanding how a person responds to a given treatment (context) for a specific malady, and whether this individual is ultimately cured, cannot happen through comparisons to another individual or by using quantitative averages computed for groups. The knowledge can only be gained by demonstrating that improvement is occurring over time and place in the individual with specificity to the interventions applied. Such a demonstration must 1. show that the individual is improving because of the specific intervention; 2. identify the measurement of the quality or speed with which improvement is or is not happening; and ultimately 3. Ascertain whether the malady was completely eliminated. An intervention (context) could not be deemed successful unless such complete results could be demonstrated. Similarly, the ubiquitous specificity of individual ⇔ context relations over the course of the human life span means that, to comprehensively and holistically understand human development, learning, and behavior across the life span, scholars need to identify the idiographic, differential, and nomothetic dimensions of a target process (e.g., Cantor et al., 2021; Lerner, 2018; Molenaar, 2004; Molenaar & Nesselroade, 2015; Rose, 2016).

Overton (2014) acknowledged that persisting difficulty existed in countering reductionists' criticism by advocates of dynamic, RDS-based developmental models. He wrote in 2014 that proponents of RDS-based models did not have sufficient methodological means to demonstrate that the models they forwarded could be rigorously and convincingly empirically interrogated. However, and as implied by the researchers using the Molenaar Manifesto and the Bornstein Specificity Principle as the frame for their studies of human development, the presence of precisely such methodological tools has been burgeoning. It is useful to briefly discuss them.

For instance, in discussing how a dynamic approach to structural equation modeling (i.e., DSEM; McNeish & Hamaker, 2020) can provide evidence of meaningful person-specific pathways of development and, as well, enable comparison of these within-person changes with the average pathway of change

for the group within which individuals are embedded, Hamaker et al. (2018, p. 820) explain that:

> *"one of the most valuable properties of intensive longitudinal data is that they provide a unique opportunity to study processes within-person as they unfold over time… To investigate the underlying dynamics of intensive longitudinal data, researchers have been borrowing techniques from other disciplines—like econometrics, physics, and engineering—where they have a long history of studying processes over time using time series analysis and dynamic systems theory. A common characteristic of these techniques is their focus on the way a preceding state of the system (e.g., person or dyad) gives rise to the subsequent state. This allows for a unique perspective on processes… and extends our more conventional approaches to intensive longitudinal data, which tend to focus on concurrent relationships between variables, rather than their dynamic interplay over time."*

At this writing, several recent studies (e.g., Abbasi-Asl et al., in press-a, in press-b; Michaelson, 2021; Yu et al., 2020, 2021, 2022a) have used DSEM to study the dynamics of attributes of academic-related skills (e.g., self-regulation, executive functioning), socioemotional functioning and character (e.g., well-being, mood, empathy), and physiological functioning (e.g., adequacy of sleep, cortisol levels).

Of course, this very brief reference to ways in which dynamic systems methods can illuminate the structure and function of person-specific relations with context–and at levels of organization that reflect the breadth of the embodied, developing person –is far from exhaustive. There are many other methods for assessing the complexity of dynamic individual ⇔ context that exist and others continue to be created (e.g., Brinberg et al., 2022; Hollenstein, 2007; Mongin et al., 2022). However, it is certainly beyond the scope of this chapter to provide tutorials about these dynamic methods. Our purpose, then, is only to point to the fact that methods to address the complexity of the embodied, holistic, and dynamic, relational development system exist and, if past is prelude, will continue to burgeon in subsequent decades.

Developmental and educational science that includes person-specific measurement and analysis enhances developmentally nurturant individual ⇔ context coactions

However, if person-specific coactions include developmentally-nurturant relations, then the probability of healthy and positive outcomes, such as thriving and learning, increases for all individuals (e.g., Cantor et al., 2021; Lerner, 1984; Lerner & Lerner, 2019). Indeed, if basic and applied research seek to move beyond only the description and explanation of developmental change among groups of individuals and, as well, to also discover ways to optimize developmental pathways through life for individuals, then understanding the bases of these developmentally-nurturant coactions will increase; scientists will learn how the theory-predicated application of such knowledge can be an important and unique additional benefit to all individuals assessed by such scholarship; of most important focus at this writing are marginalized and minoritized youth who have inequitably and unjustly developed within contexts where the proportion of adverse individual ⇔ context coactions or of insufficiently nurturant individual ⇔ individual coactions is substantially greater than developmentally-nurturant instances of such coactions (e.g., Spencer, 2024; Spencer & Dowd, 2024; Spencer et al., 2015).

Of course, individuals across their life spans should be the focus of such work. Importantly, there is voluminous literature documenting the presence of relative plasticity (the potential for systematic change and malleability) across the life span (Baltes et al., 2006; Immordino-Yang et al., 2019; Fischer & Bidell, 2006; Lerner, 1984, 2018, 2021). Nevertheless, individuals developing across the first two decades of life should be the focus of the integration of basic and applied scholarship for researchers and educators interested primarily in the K–12 span.

In sum, to understand and enhance the development of each youth in the wholeness of their specific individuality, RDS-based concepts have several important implications for research design, measurement, and data analysis that, in a useful program of scholarship, should be systematically employed. Measures that are change sensitive, valid, invariant across person, time, and place, and reliable must be used to provide data illuminating the dynamic relations between the two sides of the bidirectional arrow we have discussed throughout this chapter. In other words, measures must be created and used that are capable of

detecting changes if and when they occur because of the dynamics of mutual influence involved in coactions at all levels of the human development system, both within and across individuals, time, and place (Elder et al., 2015).

Conclusions

We believe that the future contributions of all sciences that seek to enhance the lives of humans from birth through adolescence and of course, as well, in the ensuing decades of the life course will gain greatly from framing their research and program evaluation work with ideas derived from dynamic, RDS-based models and methods. Importantly, these methods include ones that enable illumination of the ways in which idiographic fluctuations within short periods of time contribute to contemporaneous development *and* to the more directly observable qualitative and transformative changes that occur over longer periods of time, from birth to maturity. Perhaps even more important will be more discoveries of the diverse character of individual pathways and how such variation may (or may not) moderate nurturant coactions across time and place. As noted by Ram et al. (2014), a key question that we believe is vital to address in future research is: Precisely how do these fluctuations influence the transformations that reflect the qualitative changes that mark nomothetic and differential change across longer time spans and, as well, the entire course of life?

Our hope is that the documentation and dissemination of these contributions of dynamic, RDS-based models and methods and the information they carry about human growth and change will spur the growth of diverse sources of funding for basic and applied scholarship. Admittedly, theoretically-informed scholarship providing empirical evidence about the complexity of describing, explaining, and optimizing the development of holistic thriving in individual human beings is itself complex and expensive to enact. Nevertheless, only such scholarship can demonstrate how to unlock human potential and foster the adaptive development of all children and adolescents, especially youth who have been marginalized and minoritized because of living in contexts with low probabilities of experiencing developmentally-nurturant coactions. Only such scholarship can effectively replace such outcomes of development for these children, and provide for all youth, a developmental history of coactions that create vibrantly thriving individuals whose lives are marked by physical and mental health, well-being, and the acquisition of skills enabling the successful use of learning-to-learn skills across the life span.

The development of ecosystems and nurturant contexts designed for such holistic thriving will counter inequities and unjust (and undeserved) experiences (Stafford-Brizard, 2016). Such contexts will help create a citizenry that lives with and for others (Lerner et al., 2024-b). Such citizens will possess a sustained commitment to contribute in specific ways to self, family, and the institutions of civil society, social justice, and democratic civil society across the course of their lives.

References

Abbasi-Asl, R., Yu, D., Tirrell, J. M., Keces, N., Dowling, E. M., Hasse, A., Mackin, M., Olander, K., Douglas, K., Kibbedi, P, N., Wanyama, J, R., Sim, A. T. R., Lerner, J. V., & Lerner, R. M. (in press-a). Character development among Ugandan youth: A person-specific approach. *Journal of Moral Education*.

Abbasi-Asl, R., Lerner, R. M., Keces, N., Yu, D., Tirrell, J. M., Dowling, E. M., Mackin, M., Hasse, A., Olander, K., Sim, A. T. R., Lerner, J. V., Douglas, K., Kibbedi, P. N., & Wanyama, J, R. (in press-b). Measuring character among Ugandan youth: An integrative differential and idiographic approach. *International Journal of Behavioral Development*.

Allport, G. W. (1937). *Personality: A psychological interpretation*. Holt.

Allport, G. W. (1962). The general and the unique in psychological science. *Journal of Personality, 30*(3), 405–422.

Baltes, P. B., Lindenberger, U., &Staudinger, U. M. (2006). Life span theory in developmental psychology. In R. M. Lerner (Ed.), *Theoretical models of human development*. Handbook of child psychology (Vol. 1, 6th ed., pp. 569–664). Hoboken, NJ: John Wiley & Sons.

Baltes, P. B., Reese, H. W., & Nesselroade, J. R. (1977). *Life-span developmental psychology: Introduction to research methods*. Monterey, CA: Brooks/Cole.

Berkowitz, M. V., & Baer, M. C. (2024). PRIMED for character education: Deriving design principles for effective practice from empirical evidence. In M. D. Matthews and R. M. Lerner (Eds.), *Routledge International Handbooks of Multidisciplinary Perspectives on Character Development*, Volume I: *Conceptualizing and Defining Character* (pp. 104–120). New York, NY: Routledge.

Bornstein, M. H. (2017). The specificity principle in acculturation science. *Perspectives in Psychological Science, 12*(1), 3–45.

Bornstein, M. H. (2019). Fostering optimal development and averting detrimental development: Prescriptions, proscriptions, and specificity. *Applied Developmental Science, 23*(4), 340–345.

Bornstein, M. H. (2024). Parenting as panacea: Toward generational advancements of early character virtues and mature civic responsibility. In M. D. Matthews and R. M. Lerner (Eds.), *Routledge International Handbooks of Multidisciplinary Perspectives on Character Development*. Volume II: *Moderators, Threats, and Contexts* (pp. 88–115). New York, NY: Routledge.

Brinberg, M., Ram, N., Conroy, D. E., Pincus, A. L., & Gerstorf, D. (2022). Dyadic analysis and the reciprocal one-with-many model: Extending the study of interpersonal processes with intensive longitudinal data. *Psychological Methods, 27*(1), 65–81.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.

Bronfenbrenner, U. (2005). *Making human beings human: Bioecological perspectives on human development*. Thousand Oaks, CA: Sage.

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In W. Damon & R. M. Lerner (Eds.) & R. M. Lerner (Vol. Ed.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (6th ed., pp. 793–828). Wiley.

Cantor, P., & Osher, D. (Eds. (2021a). *The Science of Learning and Development: Enhancing the Lives of All Young People*. New York: Routledge.

Cantor, P., & Osher, D. (2021b). The future of the science of learning and development: Whole-child development, learning, and thriving in an era of collective adversity, disruptive change, and increasing inequality. In P. Cantor and D. Osher (Eds.), *The Science of Learning and Development: Enhancing the Lives of All Young People* (pp. 233–254). New York: Routledge.

Cantor P, Lerner R. M., Pittman K., Chase P. A., &, Gomperts, N. (2021). *Whole-child development, learning, and thriving: a dynamic systems approach*. New York: Cambridge University Press.

Card, N. A. (2017). Methodological issues in measuring the development of character. *Journal of Character Education, 13*(2), 29–45.

Chess, S., & Thomas, A. (1999). *Goodness of fit: Clinical applications from infancy through adult life*. Philadelphia, PA: Brunner/Mazel.

Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods, 25*(3), 259–270.

Collins, L.M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology, 57,* 505–528.

Deboeck, P. R., Geldhof, G. J., & Yu, D. (2023). Implications of dynamic systems for future methodology in developmental and learning science. In C. D. Lee, R. M. Lerner, V. L. Gadsden, & D. Osher (Eds.), *The science of learning and development. Review of Research in Education, Volume 47* (pp. 100–115). Sage.

Dill, H.T., & Simpson, J. T. (2024). The Templeton Philanthropies. In M. D. Matthews and R. M. Lerner (Eds.), *Routledge International Handbooks of Multidisciplinary Perspectives on Character Development*. Volume II: *Moderators, Threats, and Contexts* (pp. 535–549). New York, NY: Routledge.

Dorn, L. D., & Beltz, A. M. (2022). Puberty: Foundations, findings, and the future. In L. J. Crockett, G. Carlo, & J. E. Schulenberg (Eds.), *APA Handbook of Adolescent and Young Adult Development* (pp. 3–19). APA Publications.

Elder, G. H., Shanahan, M. J., & Jennings, J. A. (2015). Human development in time and place. In M. H. Bornstein and T. Leventhal (Eds.), *Handbook of child psychology and developmental science, vol. 4: Ecological settings and processes in developmental systems* (7th ed., pp. 6–54). Wiley.

Emmerich, W. (1968). Personality development and concepts of structure. *Child Development, 39*(3), 671–690.

Fischer, K. W., & Bidell, T. R. (2006). Dynamic development of action and thought. In R. M. Lerner (Ed.), *Theoretical models of human development. Volume 1 of Handbook of Child Psychology* (6th ed., pp. 313–399). Editors-in-chief: W. Damon & R. M. Lerner. Wiley.

Gottlieb, G. (1992). *Individual development and evolution: The genesis of novel behavior.* Oxford University Press.

Gould, S. J. (1977). *Ontogeny and phylogeny.* Cambridge, MA: Belknap Press of Harvard University Press.

Gould, S., & Vrba, E. (1982). Exaptation: A missing term in the science of form. *Paleobiology, 8,* 4–15.

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Methén, B. (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the COGITO Study, *Multivariate Behavioral Research, 53*(6), 820–841.

Heckman, J. J., Galaty, B., Tian, H. (2024). The economic approach to personality, character, and virtue. In M. D. Matthews and R. M. Lerner (Eds.), *Routledge International Handbooks of Multidisciplinary Perspectives on Character Development*, Volume I: *Conceptualizing and Defining Character* (pp. 55–103). New York, NY: Routledge.

Hirsch, J. (2004). Uniqueness, diversity, similarity, repeatability, and heritability. In C. Garcia Coll, E. Bearer, & R.M. Lerner (Eds.), *Nature and nurture: The complex interplay of genetic and environmental influences on human behavior and development* (pp. 127–138). Mahwah, N.J.: Erlbaum.

Hollenstein, T. (2007). State space grids: Analyzing dynamics across development. *International Journal of Behavioral Development, 31*(4), 384–396.

Immordino-Yang, M. H., Darling-Hammond, L., & Krone, C. R. (2019). Nurturing nature: How brain development is inherently social and emotional, and what this means for education. *Educational Psychologist, 54*(3), 185–204.

Immordino-Yang, M. H., Nasir, N. S., Cantor, P., & Yoshikawa, H. (2023). Weaving a colorful cloth: Centering education on humans' emergent developmental potentials. In C. D. Lee, R. M. Lerner, V. L. Gadsden, & D. Osher (Eds.). *The science of learning and development. Review of Research in Education,* 2023 (pp. 1–45). Sage Publications.

Immordino-Yang, M. H., & Yang, X.-F. (2017). Cultural differences in the neural correlates of social–emotional feelings: An interdisciplinary, developmental perspective. *Current Opinion in Psychology*, 17, 34–40.

Jablonka, E., & Lamb, M. (2005). *Evolution in Four Dimensions: Genetic, epigenetic, behavioral, and symbolic variation in the history of life.* Cambridge, MA: MIT Press.

Johanson, D. C., & Edey, M. A. (1981). *Lucy: The beginnings of humankind.* New York: Simon & Schuster.

Kluckhohn, C., & Murray, H. A. (1953). Personality Formation: The Determinants. In C. Kluckhohn, H. A. Murray, & D. M. Schneider (Eds.), *Personality in Nature, Society, and Culture* (2nd ed., pp. 53–69). New York, NY: Alfred A. Knopf.

Lerner, R. M. (1984). *On the nature of human plasticity.* New York: Cambridge University Press.

Lerner, R. M. (2018). *Concepts and theories of human development* (4th ed.). New York, NY: Routledge.

Lerner, R. M. (2021). *Individuals as Producers of Their Development: The Dynamics of Person Context Coactions*. New York, NY: Routledge.

Lerner, R. M., & Bornstein, M. H. (Eds.). (2021). Enriching the Study of Human Development Through the Use of the Specificity Principle: Theory, Research, and Application. *Journal of Applied Developmental Psychology, 73/74/75*.

Lerner, R. M., Bornstein, M. H., & Jervis, P. (2022). The Development of positive attributes of character: On the embodiment of specificity, holism, and self-system processes. *Human Development, 66*(1), 34–37.

Lerner, R. M., & Greenberg, G. (Eds.). (2025). *The Heredity Hoax: Challenging flawed genetic theories of human development*. New York, NY: Routledge.

Lerner, R. M., Jervis, P., & Bornstein, M. H. (2021a). Enhancing the international study of positive youth development: Process, specificity, and the sample case of character virtues. *Journal of Youth Development, 16*(2–3), 402– 422.

Lerner, R. M., King, P. E., Dowling, E. M., & Bowers, E. P. (2024-b). On being and becoming human through accompaniment and telos: Ontological convergences between Christian anthropology and the science of human development. *Journal of Psychology and Theology.* https://doi.org/10.1177/00916471241300154

Lerner, R. M., & Lerner, J. V. (2019). An idiographic approach to adolescent research: Theory, method, and application. In L. B. Hendry & M. Kloep (Eds.), *Reframing Adolescent Research* (pp. 25–38). London and New York: Routledge.

Lerner, R. M., Lerner, J. V., Murry, V. M., Smith, E. P., Bowers, E. P., Geldhof, G. J., Buckingham, M. H. (2021b). Positive youth development in 2020: Theory, research, programs, and the promotion of social justice. *Journal of Adolescent Research*, *31*(4), 1114–1134.

Lerner, R. M., & Matthews, M. D. (2024). Character development: Then, now, and next. In M. D. Matthews and R. M. Lerner (Eds.), *Routledge International Handbooks of Multidisciplinary Perspectives on Character Development, Volume II: Moderators, Threats, and Contexts* (pp. 684–704). New York, NY: Routledge.

Lerner, R. M., Yu, D., Abbasi-Asl, R., Keces, N., Gonçalves, C., Buckingham, M. H., Dowling, E. M., Tirrell, J. M., Mackin, M., Olander, K., Hasse, A., & Dunham, Y. (2024-a). Towards a dynamic, idiographic approach to describing, explaining, and enhancing the development of SEL. *Social and Emotional Learning: Research, Practice, and Policy, 4,* 1–8. https://doi.org/10.1016/j.sel.2024.100050

Li, J., & Julian, M. M. (2012). Developmental relationships as the active ingredient: A unifying working hypothesis of "what works" across intervention settings. *American Journal of Orthopsychiatry, 82*(2), 157–166. https://doi.org/10.1111/j.1939-0025.2012.01151.x

Lickliter, R. (2016). Developmental evolution. *WIREs Cogn Sci* 2016. https://doi.org/10.1002/wcs.1422

Maassen, E., D'Urso, E. D., van Assen, M. A. L. M., Nuijten, M. B., De Roover, K., & Wicherts, J. M. (2023).The Dire Disregard of Measurement Invariance Testing in Psychological Science. *Psychological Methods*. https://dx.doi.org/10.1037/met0000624

Mascolo, M. F., & Bidell, T. R. (Eds.). (2020). Handbook of integrative developmental psychology: Festschrift for Kurt W. Fischer. Routledge.

Matthews, M. D., & Lerner, R. M. (Eds.). (2024a). *Routledge International Handbooks of Multidisciplinary Perspectives on Character Development*, Volume I: *Conceptualizing and Defining Character.* New York, NY: Routledge.

Matthews, M. D., & Lerner, R. M. (Eds.). (2024b). *Routledge International Handbooks of Multidisciplinary Perspectives on Character Development*, Volume II: *Moderators, Threats, and Contexts*. New York, NY: Routledge.

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods, 25*(5), 610–635.

Michaelson, L. E., Berg, J., Boyd-Brown, M. J., Cade, W., Yu, D., Geldhof, G. J., Yang, P-J., Chase, P. A., Osher, D., & Lerner, R. M. (2021). Intraindividual fluctuations in sleep predict subsequent goal setting in adolescents. *Journal for Person-Oriented Research, 7*(2), 78–87.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*(4), 201–218.

Molenaar, P. C. M., & Nesselroade, J. R. (2014). New trends in the inductive use of relation developmental systems theory: Ergodicity, nonstationarity, and heterogeneity. In P. C. Molenaar, R. M. Lerner, and K. M. Newell (Eds.), *Handbook of Developmental Systems and Methodology.* (pp. 442–462). New York, NY: Guilford Press.

Molenaar, P. C. M., & Nesselroade, J. R. (2015). Systems methods for developmental research. In R. M. Lerner, W. F. Overton, & P. C. M. Molenaar (Eds.), *Handbook of child psychology and developmental science: Theory and method* (7th ed., Vol. 1, pp. 652–682). Wiley.

Mongin, D., Uribe, A., Cullati, S., & Courvoisier, D. S. (2022). A tutorial on ordinary differential equations in behavioral science: What does physics teach us?. *Psychological Methods*. http://dx.doi.org/10.1037/met0000517

Moore, D. S. (2002). *The dependent gene: The fallacy of nature vs. nurture*. New York: W. H. Freeman.

Moore, D. S. (2015). *The developing genome: An introduction to behavioral epigenetics*. New York: Oxford University Press.

Nesselroade, J. R., & Molenaar, P. C. M. (2010). Emphasizing intraindividual variability in the study of development over the life span. In W. F. Overton (Ed.), *Handbook of life-span development. Vol. 1: Cognition, biology, methods* (pp. 30–54). Editor-in-chief: R. M. Lerner. Hoboken: Wiley.

Noble, D. (2015). Evolution beyond neo-Darwinism: A new conceptual framework. *The Journal of Experimental Biology*, *218*, 7–13.

Nucci, L. (2017). Character: A multi-faceted developmental system. *Journal of Character Education, 13*(1), 1–16.

Nucci, L. (2019). Character: A developmental system. *Child Developmental Perspectives, 13*(2), 73–78.

Nucci, L. (2024). The Development of Morality and the Character System: Implications for the Notion of Virtue. In M. D. Matthews and R. M. Lerner (Eds.), *Routledge International Handbooks of Multidisciplinary Perspectives on Character Development*, Volume I: *Conceptualizing and Defining Character* (pp. 295–309). New York, NY: Routledge.

Overton, W. F. (2013). A New Paradigm for Developmental Science: Relationism and Relational-Developmental Systems. *Applied Developmental Science, 17*(2), 94–107.

Overton, W. F. (2014). Relational developmental systems and developmental science: A focus on methodology. In P. C. M. Molenaar, R. M. Lerner, & K. Newell (Eds.), Handbook of developmental systems theory and methodology (pp. 19–65). New York, NY: Guilford Press.

Overton, W. F. (2015). Process and relational developmental systems. In R. M. Lerner, W. F. Overton, & P. C. M. Molenaar (Eds.), *Handbook of child psychology and developmental science: Theory and Method* (7th ed., Vol. 1, pp. 9–62). Wiley.

Paus, T. (2009). Brain development. In R. M. Lerner, & L. Steinberg (Eds.), *Handbook of Adolescent Psychology* (3rd ed., pp. 95–115). Wiley.

Plomin, R. (2018). *Blueprint: How DNA makes us who we are*. London: Allen Lane.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71–90.

Raeff, C. (2016). *Exploring the dynamics of human development: An integrative approach*. New York: Oxford University Press.

Ram, N., Chow, S.-M., Bowles, R. P., Wang, L., Grimm, K., Fujita, F., Nesselroade, J. R. (2005). Examining interindividual differences in cyclicity of pleasant and unpleasant affects using spectral analysis and item response modeling. *Psychometrika, 70*(4), 773–790.

Ram, N., Conroy, D. E., Pincus, A. L., Lorek, A., Rebar, A., & Roche, M. J., Coccia, M., Morack, J., Feldman, J., & Gerstorf, D. (2014). Examining the interplay of processes across multiple time-scales: Illustration with the intraindividual study of affect, health, and interpersonal behavior (iSAHIB). *Research in Human Development, 11*(2), 142–160.

Ram, N., & Gadzke-Kopp, L. (2023). Running toward network models of individuals and their contexts. In C. D. Lee, R. M. Lerner, V. L. Gadsden, & D. Osher (Eds.), *The science of learning and development. Review of Research in Education, Volume 47* (pp. 181–184). Sage.

Rioux, C., & Little, T. D. (2020). Underused methods in developmental science to inform policy and practice. *Child Development Perspectives*, 14(2), 97–103.

Rogoff, B. (2003). *The cultural nature of human development*. New York: Oxford University Press.

Rose, L. T., Rouhani, P., & Fischer, K. W. (2013). The science of the individual. *Mind, Brain, and Education*, *7*, 152–158.

Rose, T. (2016). *The end of average: How we succeed in a world that values sameness*. Harper-Collins Publishers.

Schaie, K. W. (1965). A general model for the study of developmental problems. *Psychological Bulletin, 64,* 92–107.

Schaie, K. W., & Strother, C. R. (1968). A cross-sequential study of age changes in cognitive behavior. *Psychological Bulletin, 70*, 671–680.

Shonkoff, J. P., & Center on the Developing Child (2017). *Building a system for science-based R&D that achieves breakthrough outcomes at scale for young children facing adversity.* Cambridge, MA: Center on the Developing Child, Harvard University.

Spencer, M. B. (2006). Phenomenological variant of ecological systems theory (PVEST): A human development synthesis applicable to diverse individuals and groups. In W. Damon & R. M. Lerner (Eds.) & R. M. Lerner (Vol. Ed.), *Handbook of child psychology: Vol. 6. Theoretical models of human development* (6th ed., pp. 829–894). Wiley.

Spencer, M. B. (2024). Character virtue, social science, and leadership: Consequences of ignoring practice. In M. D. Matthews and R. M. Lerner (Eds.), *Routledge International Handbooks of Multidisciplinary Perspectives on Character Development, Volume II: Moderators, Threats, and Contexts* (pp. 633–657). New York, NY: Routledge.

Spencer, M. B., & Dowd, N. E. (2024). *Radical Brown: Keeping the promise to America's children*. Harvard Education Press.

Spencer, M. B., Swanson, D. P., & Harpalani, V. (2015). Development of the self. In M. E. Lamb (Volume Ed.), *Handbook of child psychology and developmental science*, Volume 3: *Socioemotional processes* (7th ed., pp. 750–793). Editor-in-Chief: R. M. Lerner. Wiley.

Stafford-Brizard, B. (2016). *Building Blocks For Learning: A framework for comprehensive student development*. Turnaround for Children.

Thomas, A., Chess, S., Birch, H. G., Hertzig, M. E., & Korn, S. (1963). *Behavioral individuality in early childhood*. New York: New York University Press.

Tirrell, J. M., Geldhof, G. J., King, P. E., Dowling, E. M., Sim, A. T. R., Williams, K., Iraheta, G., Lerner, J. V., & Lerner, R. M. (2019). Measuring spirituality, hope, and thriving among Salvadoran youth: Initial findings from the Compassion International Study of Positive Youth Development. *Child Youth Care Forum*, *48*(2), 241–268.

Tobach, E., & Schneirla, T. C. (1968). The biopsychology of social behavior of animals. In R. E. Cooke & S. Levin (Eds.), *Biologic basis of pediatric practice* (pp. 68–82). New York: McGraw-Hill.

Yu, D., Geldhof, G. J., Buckingham, M. H., Goncalves, C., Yang, P-J., Michaelson, L. E., Berg, J., Ni, Y., & Lerner, R. M. (2022a). "Today, I cared about how a classmate felt:" Fluctuations in empathy are linked to daily mood in adolescence. *Journal of Applied Developmental Psychology*. https://doi.org/10.1016/j.appdev.2021.101386

Yu, D., Goncalves, C., Yang, P-J., Geldhof, G. J., Michaelson, L., Ni, Y., & Lerner, R. M. (2022b). Does prior night's sleep impact next day's executive functioning? It depends on an individual's average sleep quality. *Journal for Person-Oriented Research, 8*(1), 10–23.

Yu, D., Yang, P. J., Geldhof, G. J., Tyler, C. P., Gansert, P. K., Chase, P. A., & Lerner, R. M. (2020). Exploring idiographic approaches to children's executive function performance: An intensive longitudinal study. *Journal for Person-Oriented Research, 6*(2), 73–87.

Yu, D., Yang, P-J., Michaelson, L., Geldhof, G. J., Chase, P. A., Gansert, P. K., Osher, D. M., Berg, J. K., Tyler, C. P., Goncalves, C., Park, Y., Boyd-Brown, M. J., Cade, W., Theokas, C., Cantor, P., & Lerner, R. M. (2021). Understanding child executive functioning through use of the Bornstein specificity principle. *Journal of Applied Developmental Science. 73*, 101240.

Perspectives on Socioculturally Responsive Assessment in Large-Scale Systems

Aneesha Badrinarayan, Randy E. Bennett, and Linda Darling-Hammond

Abstract

This chapter synthesizes key insights from Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy (Bennett et al., 2025), offering a cross-cutting analysis of how large-scale assessments can become more valid, equitable, and educationally meaningful when designed through a socioculturally responsive lens. The authors highlight four central themes: aligning assessment content with students' lived experiences; increasing personalization of assessments via testing processes; broadening construct definitions to reflect diverse cultural, linguistic, and epistemological perspectives; and using frameworks and tools that guide inclusive development and interpretation. Drawing on examples such as the KĀʻEO Hawaiian language immersion assessment, AP Art and Design, Smarter Balanced, and adult education programs, the chapter illustrates how sociocultural responsiveness can be integrated into both test content and administration while maintaining technical quality. It also surfaces enduring challenges, including tensions between comparability and local validity, the need for assessments to reflect community values, and the evolving role of technology in supporting adaptive and culturally grounded assessments. The authors conclude by naming critical open questions—such as whose culture is centered, how assessment use cases influence design, and how to ensure assessments both reflect and serve diverse learners—that must be addressed to advance Socioculturally Responsive Assessment (SCRA) as a transformative and sustainable paradigm for large-scale educational assessment.

Introduction

At the heart of socioculturally responsive assessments (SCRA) are goals to intentionally account for the fact that learning—and demonstration of that learning—is inherently a social and cultural activity (Nasir et al., 2020; National Academies of Science, Engineering, and Medicine, 2018), and that assessments that do not account for how students develop knowledge cannot hope to accurately surface and communicate what a diverse range of students knows and can do. This fact influences both what we assess as well as how we surface evidence of learning. SCRA (Bennett 2023) can support culturally responsive, relevant, and sustaining teaching and learning through two focal strategies: (1) instruments that are themselves designed to be socioculturally responsive, and (2) instruments that incentivize and support culturally responsive, relevant, and sustaining pedagogical practices at various levels within the system (Badrinarayan et al., 2025). Identifying features of assessment systems that can meet one or both of these goals requires understanding major findings and advances across bodies of work focusing on the most effective teaching, learning, and assessment approaches for specific groups of students, and extending those learnings to assessment design, development, and implementation.

In compiling the recent edited volume *Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy* (Bennett et al. 2025d), we sought to synthesize a range of leading perspectives on approaches, technical considerations, and instrument and process designs intended to serve a common purpose: the development of assessments that are responsive to the unique and shared social, cultural, and linguistic experiences that shape how people learn and how they make what they know and can do visible. The resulting volume intentionally explores the diversity of ideas related to the conceptualization of SCRA, technical approaches for contending with culture and language at scale, working examples and policy considerations of systems that have centered culture and language in their design and implementation, and ongoing research to provide empirical evidence for how the attention to sociocultural factors contributes to learning and performance so learners make what they know and can do visible on large-scale assessment instruments. This chapter offers insights drawn from across the range of scholarship present in the volume to better characterize SCRA at scale, focused on emergent themes related to:

1. Relevance of test content
2. Personalization via testing processes
3. Broadening of construct definition
4. Assessment development frameworks for SCRA

Taken together, the ideas reflected in this volume provide important considerations for large-scale systems intended to support learning as a primary goal. It should be noted that while many scholars and practitioners have considered how to make assessments more culturally responsive (see Bennett, 2025 for a detailed review), this chapter focuses on the specific insights offered by the authors within our volume.

Theme One: The Importance of Relevance of Test Content to Students' Lived Experiences

Across many different conceptions, socioculturally responsive assessment design is predicated on the idea that how examinees interact with an assessment is not a fixed feature of assessment design, and instead is inextricably linked to the social, cultural, and linguistic contexts within which learning and performance happen. A goal, then, of SCRA is to bring assessment instruments into better alignment with examinees' experiences through various approaches to increasing relevance and personalization (Bennett 2023, 2024, 2025). In assessments that are closer to the classroom, relevance and personalization are often achieved through deep relationships with the learners taking the assessment. For example, a teacher may account for their students' home languages and shared classroom experiences by adapting a unit assessment to better match their students' specific lived experiences; a different teacher may select particular texts she knows have relevance to interests and identities present in her classroom when designing an assessment of reading comprehension (Ebe, 2025). Such approaches to personalization and relevance are often considered best practice for responsive teaching when done in the classroom, but become infinitely more challenging when similar principles and expectations are extended to large-scale assessments that (1) operate across many different—and dynamic—learning contexts, student groups, and student experiences (2) are developed by assessment designers who do not have relationships with learners/examinees, and (3) often have fewer degrees of freedom for local adaptation by students and/or teachers. Authors in the volume

explore the relationship between students' funds of knowledge, test relevance, and performance through many different lenses, including disciplinary perspectives, approaches that attend to cultural ways of knowing, and ideas that account for relevance by broadening our conceptions of measurement targets.

Relevance Within Disciplinary Contexts: Insights From Reading Assessments

Reading assessments have been an area of focus for SCRA because of the considerable evidence that students' experience with and background knowledge of a given topic influences how they understand and make meaning of related texts. Text selection for large-scale assessments is a promising and important direction for content-based approaches to personalizing assessment instruments. Most large-scale reading assessments focus on common, provided texts as the basis for measuring students' abilities, and many qualitative and quantitative features of texts (e.g., text complexity, text types) can be accounted for while still allowing for more socioculturally grounded decisions about the nature of the texts students interact with on assessments. Ebe (2025) offers *The Cultural Relevance Rubric* as one way to identify whether texts are culturally relevant to the students engaging with them. The rubric, which is described in detail later in this chapter, asks students to rate texts on a series of factors related to age, geography, language use, and types of activities discussed, to determine how close the texts are to given students' experiences. The rubric gives students agency in determining the relevance of texts, and Ebe suggests that the rubric could be used either as a method for selecting the most appropriate text on a given instrument or be reported alongside reading scores to provide important context for interpreting student performance and planning next steps.

Similarly, Skerrett and colleagues (2025) discuss how large-scale reading assessments can better account for the social nature of reading, students' funds of knowledge, and the range of student experiences with texts and topics. Drawing on the development of the 2026 National Assessment of Educational Progress Reading Framework, Skerrett et al. suggest several strategies for socioculturally responsive reading assessments, including:

- Multimodal knowledge scaffolds that leverage video, audio, and visual supports to support background knowledge students may need to access and make sense of the texts present on the exam;
- Allowing students to respond in home languages; and
- Embedded contextual probes (similar in many ways to the questions Ebe [2025] proposes in the Cultural Relevance Rubric) that can help assess student interest, motivation, and familiarity with texts.

Wang and colleagues (2025) offer empirical evidence to support the idea that using culturally relevant texts affects students' reading comprehension performance on large-scale assessments. They explore how Black and non-Black student performance on reading assessments changed based on the relative familiarity and relevance of the texts used on each form. They found that Black students had significantly less background knowledge on two of the three topics—immigration and ecosystems—but had similar levels of background knowledge on the third topic, the Harlem Renaissance. Their performance mirrored this finding: Black students spent more time on task and performed better on the Harlem Renaissance form than on the other two forms. Importantly, using the form with higher relevance and familiarity to Black students did not adversely affect non-Black students' performance, nor did the high-relevance form demonstrate any threats to psychometric quality. Taken together, these findings are consistent with calls for more culturally attentive considerations when choosing texts, and suggests that text selection can be a critical way to personalize and make more relevant large-scale assessments when those selections are made with explicit considerations for the experiences of the diverse examinees intended to be taking the assessment.

Relevance to Specific Cultural and Linguistic Ways of Knowing.
A vital component of many conceptualizations of SCRA is the concept of cultural validity—the extent to which assessments account for the cultural, linguistic, and social repertoires students draw upon when interpreting and responding to disciplinary tasks (Solano-Flores & Ruiz-Primo, 2025; Nelson-Barber & Trumbull, 2025; Solano-Flores & Nelson-Barber, 2001). Cultural validity concerns how these repertoires shape meaning-making and influence not only performance but also the validity of inferences drawn about student understanding.

Nelson-Barber and Trumbull (2025), drawing on work with Indigenous communities, argue that cultural validity is as essential as construct validity: disregarding students' cultural and linguistic contexts during assessment development and score interpretation introduces threats to overall validity. For example, for Indigenous students—including American Indian, Alaska Native, Native Hawaiian, and other Pacific Islander groups—histories, governance structures, belief systems, heritage languages, cultural values, and community-based practices are distinct from Western, Eurocentric norms. When assessments ignore these lived realities, they may:

- Present information in unfamiliar or incongruent ways that create barriers to comprehension;
- Require response modes that conflict with students' typical ways of demonstrating understanding;
- Apply narrow success criteria that devalue culturally rooted ways of knowing and sensemaking; and
- Reinforce experiences of dehumanization, marginalization, or erasure, which can undermine motivation and perseverance.

In all such cases, the result is the same: student performance no longer reflects what students know and can do, but rather their ability to navigate unfamiliar or invalidating assessment structures.

While these issues have been examined carefully in the assessment of Indigenous students, they also extend to many other populations whose cultural and linguistic experiences are underrepresented in mainstream testing assumptions. These groups include multilingual learners; students with disabilities; students from urban, rural, or remote communities; students from low socioeconomic backgrounds; and immigrant or migrant youth, as well as students encountering assessments in countries other than the United States. For example, Raji and Baidoo-Anu (2025) examined the cultural relevance of test items on the West African Senior School Certificate Examination, a high-stakes university entrance exam used in Ghana and Nigeria. They found that some items were culturally problematic for the test-takers expected to engage in the exam: for example, stimulus information was sometimes culturally disconnected from West African students' experiences, and test language was in some instances unnecessarily complex. Based on their analysis, they

raise concerns that the exam may potentially unfairly limit or deny postsecondary opportunities for students.

Many assessment developers find balancing the fundamentally different world views required to enhance cultural validity nearly insurmountable in large-scale assessment design. However, Nelson-Barber and Trumbull (2025), Solano-Flores and Ruiz-Primo (2025), and Raji and Baidoo-Anu (2025) recommend some concrete steps that assessment developers can embed within assessment content development to strengthen cultural validity and ensure assessments are experienced as more relevant and valid measures of student understanding and ability:

- Define a clear theory of action connecting assessment content, student experiences, and intended outcomes;
- Consider how students' cultural experiences and ways of knowing influence interpretation of test content and response strategies—do not assume any particular understanding or premise is "universally" understood unless there is strong evidence to suggest all test-takers will have had sufficient experiences to have a shared understanding; this consideration might include some degree of population mapping to better understand the cultural and linguistic experiences and assets of the students to be engaged in the assessment;
- Right-size linguistic and semiotic complexity so that all students can access and engage deeply with assessment tasks, including through multimodal scaffolds and contextual cues that serve to enhance engagement without limiting opportunities for rich meaning-making;
- Broaden interpretations of what counts as a "correct" or valid response, accounting for diverse ways and modes of sensemaking;
- Include culturally and linguistically representative students throughout the assessment development process (e.g., through interest surveys, cognitive labs, co-design efforts); and
- Provide opportunities for open-ended responses that enable students to show their thinking in rich, authentic ways. These opportunities should include allowing for multiple modes of expression and authentic engagement with cultural funds of knowledge, community-based knowledge systems, and diverse perspectives and values (e.g., as part of selecting claims or making meaning of presented stimuli).

Creating relevant large-scale assessments fundamentally requires that test developers generate content by making some assumptions about students—assumptions can range from extrapolating insights from deep and intentional student engagement to a broader set of students, to making assumptions that certain aspects of students' backgrounds (e.g., race, ethnicity) will confer experiences or perspectives to all students who share those elements of identity. Authors in the volume repeatedly caution against essentializing communities—reducing complex, dynamic cultures to fixed traits or stereotypes. Instead, assessment design must be grounded in deep understanding of the specific histories, languages, and epistemologies of communities. For example, in the case of Indigenous students, this grounding might include recognizing their status as sovereign nations and respecting their self-determined educational and cultural frameworks (Nelson-Barber & Trumbull, 2025), as well as intentionally and explicitly contending with these factors when designing large-scale assessments for those students.

Theme Two: Right-Sizing Standardization of Testing Processes to Increase Personalization

Another approach to bringing assessment experiences into better alignment with students' social, cultural, and linguistic experiences is allowing for personalization or user-specific adaptations in ways students interact with assessments. Sireci and colleagues (2025) describe a framework for UNDERSTANDardization in which decisions about what aspects of test administration are standardized are driven by learner-centered approaches that seek to better understand how different groups of students might understand and interact with the test, rather than rigid models of standardization that emphasize uniform testing conditions as a primary way to ensure fairness and validity of score interpretation. Sireci et al. (2025) discuss UNDERSTANDardization in the context of two adult-centered assessments—the Adult Skills Assessment Program (ASAP) and the English Test for Adults (ETA)—both of which are intended to be administered to adults at scale. By leveraging literature reviews and focus groups to understand user experiences of both the examinees and the educators and employers who might use the resulting data, researchers identified several ways for being more flexible in the assessment administration. Such methods as translanguaging options and real-time translations of instructions, choice in task format, and adaptive pacing could potentially improve score validity and the utility of the assessments.

Sireci and colleagues' UNDERSTANDardization framework provides a useful model for examining administration-related approaches to SCRA. In the volume, authors describe two major approaches to this kind of personalization in K–12 assessment systems: (1) adaptations and accommodations that allow learners to better access the *same* content, and (2) allowing for student agency in deciding how students will demonstrate evidence of progress, proficiency, and mastery relative to common constructs via different content. For example, Michel and Shyyan (2025) describe the accessibility supports within the Smarter Balanced assessment system (SBAC) as a set of strategies for personalizing the testing experience such that all learners—particularly students with disabilities and multilingual learners—can engage meaningfully with the same test content. Smarter Balanced assessments are built around relatively narrow constructs tied to states' math and English-Language Arts standards and are designed to be administered across diverse state contexts and student backgrounds. The assessments' primary goal is to generate evidence of student progress toward academic standards for federal accountability purposes within each state, which requires a certain amount of rigidity in content and administration. Despite these requirements, the assessment system incorporates universal design features, such as language supports and alternative response modes, that allow students to access and demonstrate targeted knowledge and skills. This strategy of targeted personalization within a standardized system offers one potential pathway for making large-scale assessments more socioculturally responsive.

Smarter Balanced provides multiple levels of support for personalization, including:

- Universal tools that are available to all students (e.g., digital notepad), allowing them to customize their testing experience to better account for how they make meaning and access content;
- Designated supports that are available to students with educator-identified needs (e.g., glossaries in 13 languages) to support specific kinds of engagement; and
- Accommodations for individual students with documented IEP or 504 plans (e.g., braille versions).

These supports are designed to ensure that each student engaging with the assessment has the best opportunity to show what they know and can do. The intention is to *enhance* construct comparability, rather than detract from it by ensuring that differences in performance reflect differences in skill, not access. While these supports introduce some variability in test administration procedures, ranging from minor (e.g., notepad use) to more substantial (e.g., braille), they increase fairness by enabling students to more accurately demonstrate competency. This approach illustrates how standardized assessment systems can integrate flexibility to better serve a diverse population without compromising measurement integrity.

In a distinct disciplinary context and serving different purposes, Escoffery and colleagues (2025) present the Advanced Placement (AP) Art and Design examination as a compelling model for integrating SCRA principles into large-scale assessment design. This through-course portfolio examination operates at the intersection of formative and summative assessment, targeting a set of constructs that readily support flexible interpretation in terms of topic, content, and medium of demonstration, such as sustained inquiry; practice, experimentation, and revision; synthesis of materials, processes, and ideas; and creative artmaking. What distinguishes AP Art and Design is how it approaches standardization: Students are afforded significant autonomy to engage in artistic production across diverse media, enabling them to draw upon their cultural identities, personal experiences, and individual interests as integral components of demonstrating proficiency. This design feature explicitly positions the student's sociocultural background as a resource rather than a variable to be controlled. What is more rigorously standardized is the evaluation criteria and scoring process. A clearly articulated construct definition, operationalized through a detailed rubric and implemented by highly trained art educators, ensures construct comparability across an inherently diverse array of student submissions—from mat-board constructions to paintings to interactive installations. While the student work varies substantially in form and content, score comparability is maintained through robust rater training, the use of shared interpretive standards, and a calibrated scoring process. This approach illustrates a deliberate and productive tradeoff: by foregrounding construct clarity and scoring consistency, the assessment system supports valid and reliable credit and placement decisions while simultaneously honoring student agency in form, expression, and cultural perspective.

Indeed, principled design for student agency like that seen in AP Art and Design, in which students can directly play a role in appropriately personalizing assessments by making choices grounded in their own lived experiences, is a promising direction for assessments that seek to be authentically responsive to students without the harms of developer-based assumptions about what is most relevant or meaningful to learners and their performance. Many chapters in the volume touch on aspects of student agency—for example, Ebe's (2025) Cultural Relevance Rubric relies upon student-determined relevance, and Badrinarayan and Darling-Hammond (2025) describe several national and international assessment systems that include elements of student choice and agency in determining the tasks and topics with which individuals will engage (e.g., through AP and International Bacalaureate assessments).

Theme Three: Expanding Construct Definitions to Account for Social, Cultural, and Linguistic Contributions to Learning and Performance

Across arguments for increasing the relevance and personalization of assessments for specific learners, there has been an implicit—and at times explicit—call to rethink both what assessments are measuring and how those measurements are interpreted. This includes calls to:

1. Expand the range of assessment constructs to better reflect how learning and development actually occur;
2. Incorporate culturally and community-specific priorities and goals for student learning; and
3. Develop more trustworthy and inclusive measures of what students know and can do, even within narrowly defined domains.

Together, these shifts aim to ensure that assessments more accurately reflect the full breadth of student learning and experience.

Lee (2025) contends that the Science of Learning and Development (SoLD)—which synthesizes interdisciplinary research from human development, psychology, neuroscience, and the learning sciences to offer a comprehensive understanding of the diverse factors that shape learning across the lifespan—requires a fundamental rethinking of the aims and design of educational assessments. That rethinking has the potential to yield more actionable, equitable, and ecologically valid insights.

Lee argues that persistent disparities in assessment outcomes—by race/ethnicity, gender, and socioeconomic status—stem not only from inequitable opportunities to learn but also from the limitations of existing assessment systems. Current summative, interim, and formative approaches fail to capture the full range of influences on how people learn and demonstrate knowledge. These influences include cultural identity, perceptions of the task and setting, emotional salience, epistemological beliefs, mindsets, and self-efficacy.

Moreover, Lee critiques the dominance of narrow, Eurocentric definitions of disciplinary knowledge in U.S. education, which constrain both teaching and assessment. She argues that expanding our conceptions of what counts as knowledge—in ways that are more culturally and contextually responsive—could reduce disparities, foster a more holistic view of learners' capabilities, and better recognize the strengths and knowledge systems students from historically marginalized communities bring to school.

Lee's challenge to narrow definitions of disciplinary knowledge in current large-scale assessment systems is echoed in many other chapters in the volume. For example, Welch and Dunbar (2025) examine opportunities for integrating SCRA into federally mandated state assessments. Based on their analysis, they argue that current interpretations of alignment in both item development and item/test evaluation (e.g., for purposes of federal peer review) privilege overly narrow conceptions of what it looks like to demonstrate performance relative to established standards. They suggest that the underlying reason for these narrow conceptions are at least two-fold. First is that the standards themselves were often not developed with cultural relevance in mind—expectations for student performance are often imbued with White, Western-dominant perspectives and worldviews that then are operationalized in assessments accordingly (e.g., language standards that prioritize standard written English, use of terms like "effective" and "appropriate" which can marginalize students with diverse linguistic repertoires and serve to reinforce dominant structures in terms of whose culture is valued). Second, current conceptions of assessment alignment focus on very narrow interpretations of item-standard mapping and matching; if instead alignment were considered more holistically and inclusively, that consideration might allow for culture and lived experience to play a more meaningful role in how items surface—and students demonstrate—progress toward the goals being measured on an assessment, with

the result that large-scale math and ELA assessments would be better positioned to support student progress.

Kukea-Shultz and Englert (2025) describe how ideas related to reimagining what is assessed on large-scale, federally mandated assessments are factored into a culturally-specific operational assessment: the Kaiapuni Assessment of Educational Outcomes (KĀʻEO), the accountability assessment for students attending Hawaiian language immersion programs. Developing KĀʻEO required explicitly countering the standard, monocultural worldviews that govern most large-scale assessment development processes because central to KĀʻEO's purpose is the reclamation of Hawaiian culture and language; this reclamation required defining constructs in ways that are responsive to the local community's language, culture, and information needs while still meeting federal requirements for showing progress in mathematics, language arts, and science. Kukea-Shultz and Englert (2025) describe specific ways in which the measurement targets for KĀʻEO were culturally defined, including:

- Developing culturally and linguistically specific student learning outcomes that align with the intent and goals of Hawaiʻi's state standards in math, language arts, and science (Common Core State Standards and the Next Generation Science Standards) but reflect Hawaiian linguistic and epistemological priorities;
- Extending measurement goals beyond state standards to include culturally-specific forms of knowledge (e.g., use of metaphoric language) that are not generally prioritized in language arts assessments;
- Using the value of Hawaiian language, knowledge, and culture as a consistent lens throughout the assessment development process.

KĀʻEO also relies upon a deeply collaborative and relational approach, placing educators and community members within the Kaiapuni system as decision-makers throughout the assessment development and validation process. For example, teachers, families, and community members are directly engaged as part of the team establishing the student learning outcomes and goals of the assessment; teacher judgments of student performance are used as part of the process for validating test scores; and culturally relevant opportunity-to-learn measures are included as part of reporting efforts to facilitate collaborative, action-oriented meaning-making around student performance. Kukea-Shultz and Englert (2025) posit community

validity—reflective of the intentional processes governing KĀʻEO's development and validation—as a framework for assessment development that positions assessment as an activity that serves communities as a first-order principle.

Many scholars (e.g., Lee 2025, Nelson-Barber and Trumbull 2025, Kukea-Schultz and Englert 2025) emphasize the need to disrupt the dominance of White, Eurocentric cultural norms that shape not only how assessments are designed, but also how students experience schooling more broadly. Zandvakili and Gordon (2025) highlight what is often the elephant in the room—that European and American cultural frameworks continue to govern the operation of much of the developed world. They argue that while education must necessarily contend with these dominant norms, it must also intentionally accommodate the diverse cultural backgrounds, identities, and experiences of learners.

To do so, they propose designing assessments that cultivate, recognize, and incentivize a broader range of competencies—competencies that emerge through students' negotiation of their own cultural frameworks within an increasingly diverse and interconnected world. The authors describe five core competencies that would serve to equip learners to thrive in both dominant and marginalized systems while contributing to inclusive educational environments that value and validate a wide range of intelligence and knowledge:

- Accession: Embracing diverse perspectives and cultural knowledge;
- Accommodation: Adjusting behavior and mindset in response to different environments;
- Adaptation: Navigating new or evolving situations with flexibility;
- Adjustment: Fine-tuning strategies in response to feedback and challenges; and
- Agility: Thinking creatively and critically to solve complex problems.

Centering these competencies requires a new approach to assessment. Zandvakili and Gordon (2025) suggest that rather than relying on passive instruments rooted in a single cultural worldview, systems could center agentic assessments that are responsive to students' individual contexts while also actively inviting learners to apply their cultural and linguistic assets to engage with sociocultural perspectives different from their own. For example, culturally universal probes and agency-oriented probes might ask students to connect their learning to personal, cultural,

and community-based experiences and problem-solving of their own choosing, while critical thinking probes might encourage students to explicitly take multiple perspectives, practice empathy, and engage in deep analysis.

Across testing contexts, many of the book's chapter authors contend that transforming assessment requires a fundamental shift in what we value, how we define competence, and whose knowledge is recognized. By bridging definitions of what we measure—and what we value—with the full diversity of learner experiences, they suggest that assessment systems can become more inclusive, relevant, and empowering.

Theme Four: Assessment Development Processes that Account for Sociocultural Goals

A fourth theme that runs through the book is the need for and use of frameworks and tools to guide assessment development and score interpretation in ways that align with the conceptual and evidence-based recommendations for SCRA. Frameworks have long been used with standardized tests for such purposes. Classical test theory (Gulliksen, 1950), item response theory (IRT; Lord, 1980), evidence-centered design (ECD; Mislevy et al., 2003), and the argument-based approach to validation (Kane, 1992) are widely used examples. Frameworks are important because, in the best case, they offer principled approaches for thinking about a problem and taking action to address it.

SCRA presents no shortage of problems, including ones related to assessment design, development, analysis, and interpretation. Key to SCRA design is taking account of examinee sociocultural characteristics to allow individuals to demonstrate better what they know and can do. Sato (2025) offers an approach to design that focuses on deeper levels of culture, with the intention of accounting for those factors (e.g., values, norms, beliefs, language, social structure/dynamics, milieu) that affect students' meaning making and their representations of knowledge. The chapter presents a *sociocultural dimensions matrix* describing personal orientations. The matrix should be of use in designing more inclusive measurement targets, tasks, and scoring rules, as well as for guiding interpretations of diverse student performance.

The matrix is organized around three sociocultural dimensions that can influence an examinee's comprehension and, hence, their expression of knowledge: Social Relationships/Orientation; Epistemological Beliefs/Cognitive Patterns; and Communication Patterns. The social relationships/orientation dimension reflects such propensities as individualistic vs. collectivist and nurturing vs. challenging patterns of behavior, whereas the epistemological beliefs/cognitive patterns dimension concerns tendencies toward analytic vs. holistic and random vs. sequential thought. Five broad communication patterns are delineated: English, Romance (e.g., Spanish, French, Italian), Semitic (Hebrew, Arabic), Asian (Japanese, Vietnamese, Mandarin, Cantonese), and Russian. As noted in the matrix, the logical structure found in English tends toward the deductive presentation of information, with ideas related in an orderly sequence. In contrast, romance languages have a logical structure that is more likely to engender lines of thought that sometimes pursue complex digressions. Semitic logical structures may produce parallel lines of development, including tangential information, whereas Asian structures may encourage circular reasoning by indirection. Finally, Russian communication patterns typically entail one or more lines of development.

As Sato notes, the matrix suggests—and research supports—the contention that individuals with particular orientations process information differently. For example, the orientation of European Americans toward analytic ways of thinking may result in taxonomic reasoning whereby objects are categorized conceptually based on shared attributes. In contrast, Chinese who tend toward collectivist and holistic orientations, are more inclined toward relational reasoning, which may lead to grouping items on common functions. Glick (as cited in Greenfield, 1997) relates the consequences of such differences in orientation for ability test performance using a categorization task with Liberian subjects. Repeated trials across multiple examinees resulted in functional groupings (e.g., potato and knife) because that was what a "wise man" would do, rather than the expected separate conceptual groupings of foods and of tools. When asked how a fool would organize the objects, the participants quickly produced the expected conceptual groupings. Clearly, the examinees' and examiners' notions of intelligent behavior were culturally determined—and opposite one another. In short, sociocultural orientations matter. Sato's (2025) matrix offers a framework for understanding them better and acting upon that understanding, especially for purposes of assessment design and interpretation.

In her chapter, Ebe (2025) continues the concern with more effectively accounting for sociocultural orientation in assessment. As noted above, her focus is upon the cultural relevance of text in reading passages. In line with Sato's (2025) chapter, as well as with much reading research (Lee, 2025; Wang et al., 2025), Ebe's premise is that comprehension depends critically upon relevant prior knowledge. We more quickly and completely understand text that draws upon what we already know. Moreover, what we know is culturally shaped, as Glick's experience (cited in Greenfield, 1997) so memorably attests. Thus, knowing the cultural relevance of text should help in SCRA design and score interpretation. As described above, Ebe (2025) offers a tool to do just that. The cultural relevance rubric consists of eight questions, each of which is rated on a 1–4 Likert scale. The questions concern the characters, setting, and experiences described in any given text. The ratings, typically done by the student, are intended to index the proximity of the text to the student's lived experience.

Several previously published studies support the utility of the rubric. In one study, Ebe (as cited in Ebe, 2025) asked 3rd-grade emergent bilingual students to read and retell two stories from a commercial reading kit that identified the stories as being at the same level. After reading each story, students rated its relevance. The recordings of each retelling were then analyzed using miscue analysis to identify how well readers were using semantic, syntactic, and graphophonic cues, along with background knowledge to comprehend the text. All students were more proficient reading the story they rated as more relevant. Retellings were also more accurate, detailed, and complete for the more relevant story.

Ebe (2025) suggests two potential uses for the rubric. One use might be as a guide to text selection, with student ratings collected as part of a text-evaluation phase in examination development. The desired result would be to locate relevant texts for each numerically significant demographic group. A second use might be for scoring and interpretation. In this use, test-takers would rate the passages they read. Along with a conventional score, scores that weighted items according to passage relevance could also be computed, thereby giving test users a sense of how comprehension varied with perceived relevance for each student.

A general framework for guiding the development of SCRAs is offered in the chapter written by Badrinarayan and Darling-Hammond (2025). The framework was derived from a review of state and national attempts to account for sociocultural factors in large-scale assessment systems. The reviewed systems were of four types: (1) assessments constructed for a specific community's language and culture (e.g., students attending Hawaiian language immersion programs); (2) assessments that are embedded in High Quality Instructional Materials developed to be culturally responsive (e.g., Washington State's use of OpenSciEd); (3) assessments whose items should appeal to a wide array of cultural identities because they cross such factors as race/ethnicity, age, gender, language, immigration experience, disability, and geography (e.g., NAEP 2028 Science Framework); and (4) assessments that allow students to choose among tested subjects and/or problems (e.g., Finnish matriculation exam), as well as design their own problems (e.g., AP Research), enabling students to bring their interests, prior knowledge, and lived experience to bear.

Badrinarayan and Darling-Hammond's framework consists of five features intended to be used as "… a heuristic for defining the potential landscape of culturally conscious assessment systems at scale" (p. 362). The first feature is to *Emphasize authenticity, agency, and decision making.* Consistent with this feature are real-world tasks that call upon disciplinary modes of inquiry, allow students choice in what and how to inquire, and permit demonstration of competency through multiple avenues. The second feature is to *Create tasks that are relevant and meaningful to specific communities.* This feature is instantiated through assessment problems that are deemed important to communities, call upon students to generate solutions aimed at positive impact, and ask students to apply their disciplinary knowledge to address issues, keeping in mind social, economic, environmental, and political concerns. The third feature is to *Center asset-based narratives of minoritized people and communities, bringing diversity to those positioned as knowers and doers.* Assessment tasks consistent with this feature include ones that represent non-dominant people as role models in the discipline and world, avoiding superficial or stereotypic depictions. Fourth is to *Emphasize dynamic relationships to cultural relevance and perspective taking.* In keeping with this feature, tasks should be designed to simultaneously reflect the cultural experiences of some students, while acting as cultural learning opportunities for other students. Additionally, task design should encourage students to bring their own ideas, experiences, and perspectives to sense-making about the phenomenon under

study. The last framework feature is to *Engage positive, productive affect and effort.* The intention behind this feature is to build tasks around interesting, compelling phenomena that cause students to engage and persevere. Collaboration is cited as another mechanism for fostering a positive context for problem solving.

Sato's (2025), Ebe's (2025), and Badrinarayan and Darling-Hammond's (2025) chapters principally focus on frameworks and tools for SCRA development. In contrast, Moses (2025) deals primarily with issues of analysis and score interpretation. Moses contrasts and compares sociocultural theories of learning and development with the practices used in large-scale assessment for test scoring and linking. He notes that tests developed from a socioculturally responsive perspective may vary across examinees by, among other things, presenting different questions or posing them in ways that suit test takers' backgrounds. Large-scale assessment practices, in contrast, suggest that the maintenance of scores would ideally require that the tests be designed to the same constructs and specifications rather than adjusted to the backgrounds of different test takers. Moses depicts the challenge as a tradeoff between comparability across examinees vs. validity for a particular use and/or examinee group. He suggests a framework consisting of three ways in which this *local validity* vs. *broad comparability* tension might be resolved.

The first possibility is to standardize to the most appropriate group or construct, where the construct and assessment are engineered from the outset to be responsive to that group (e.g., KA'EO; Kūkea-Shultz & Englert, 2025). To facilitate score comparability, subsequent test forms are created, given, and scored following standardized procedures. Under this possibility, score interpretations are similar for all examinees, within and across test forms. Moses' second possibility is to keep scores inferences local, meaning particular to the individual or to that subset of individuals taking essentially the same assessment under similar conditions. Comparability across the entire group is restricted but validity for individuals or specific groups may be enhanced. The last possibility is to expand the construct definition to account for the variation resulting from a responsive test. For example, on an assessment that allows examinees to create problems (e.g., AP Research), their design choices will, implicitly or explicitly, become part of what is measured and, therefore, what examinees are being compared upon. By virtue of allowing problem creation, each examinee takes a different test, each test intended to measure the same high-level construct (e.g., the ability to design, conduct, and

defend a study). The expanded construct definition, a rubric to connect disparate examinee performances to that definition, and rater training/monitoring processes then become mechanisms for generating scores that are both roughly comparable and locally valid.

Like Moses, Mislevy et al. (2025) deal with the issues posed by SCRA for analysis and score interpretation, but additionally with the implications of SCRA for development. These authors propose a lens that connects the logical assessment-argument structures of Evidence-Centered Design (Mislevy et al., 2003) with the sociocultural aspects of tasks (e.g., aspects that are construct essential, related to sociocultural background, ancillary, enabling or restricting, genre specific). That combination allows for analyzing the relations among tasks, students, purposes, and inferences to facilitate assessment design decisions. The combination also allows for a better understanding of SCRA's potential effects on score validity vs. comparability.

Mislevy et al. make the useful distinction between *data comparability* and *construct comparability*. Data comparability results when all examinees take the same assessment under the same conditions such that the method of assessment is common. Data comparability does not necessarily imply construct comparability, although it may. In a reasonably homogenous population, using a common method will produce both data and construct comparability. However, when the examinee population is highly heterogenous, examinees will bring different understandings to the assessment that may introduce irrelevant difficulty, thereby compromising construct comparability (see Glick, as cited in Greenfield, 1997). By customizing the assessment to examinee groups or individuals, SCRA works to achieve construct comparability, though at the expense of lower data comparability.

Dealing with the consequence of lower data comparability calls for what Mislevy et al. deem a "rectification argument," in essence a mechanism for placing scores from disparate assessments on the same scale. Different types of rectification may be appropriate depending upon the situation, context, desired inference, populations, and differences across the assessments given to individuals. When parallel forms are administered with common anchors or common persons (e.g., as for the SAT), rectification can be achieved through standard equating procedures. When the assessments are not parallel but the respective populations are sufficiently similar that students can also take common anchors, concordances may be established that allow more limited types of inference (e.g., for groups

rather than individuals). In cases where both the assessments and the examinee populations are different, as is true for many SCRAs, score rectification can be approximated through other mechanisms, including expert judgment, the use of common higher-level rubrics, and theories that can be used to map different assessments to the same underlying framework (e.g., learning progressions).

Conclusion and Next Steps for SCRA

This chapter draws together a broad and evolving set of ideas, tools, and practices that are redefining what it means for large-scale assessments to be valid, fair, and educationally meaningful in increasingly diverse learning contexts. Across the many contributions to the source volume *Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy* (Bennett, Darling-Hammond, & Badrinarayan, 2025b), one message emerges with clarity: SCRA requires more than technical improvements or inclusive messaging—it demands a transformational shift in how we conceptualize, design, implement, and interpret educational assessments at scale.

This shift begins by re-centering the foundational assumption that learning is inseparable from culture, identity, and context, and that assessment must reflect this reality. The chapters in the volume demonstrate how assessment systems grounded in sociocultural responsiveness can be designed to recognize multiple ways of knowing, support meaningful engagement with academic content, and honor the full range of students' cultural and linguistic repertoires. Whether by increasing the relevance of content, personalizing assessment processes, or broadening construct definitions, the work described here challenges dominant paradigms that have historically marginalized non-dominant learners.

Importantly, contributors show that technical quality and sociocultural responsiveness are not mutually exclusive. Through examples such as the KĀʻEO assessment in Hawaiʻi, the AP Art and Design portfolio, and Smarter Balanced's universal design features, the chapters illustrate how construct comparability can be preserved or even strengthened when design decisions are guided by community-informed theories of action, clearly articulated validity arguments, and inclusive development practices. This consideration includes not only attention to item and task development, but also to scoring practices, data use, and communication of results—all of which must be revisited to better serve students and communities.

To that end, the volume also highlights the value of new frameworks and analytical tools—including cultural relevance rubrics, sociocultural design matrices, and approaches to UNDERSTANDardization—that help operationalize SCRA in practical terms. These tools underscore the need to decenter monolithic conceptions of standardization and instead embrace models that prioritize construct validity over procedural uniformity, allowing for more nuanced, community-centered, and asset-based approaches to large-scale measurement.

While the field continues to wrestle with trade-offs between comparability and responsiveness, efficiency and authenticity, and alignment to standards and cultural relevance, the authors in this volume make clear that we must move beyond binary thinking. As Moses, Sato, Kukea-Shultz and Englert, and others argue, validity arguments must be both technically sound and socially credible, grounded not just in statistical evidence, but in the lived realities and values of the communities assessments are intended to serve.

Whereas this chapter highlights common themes across a wide range of ideas in the book, there are many open questions with regard to SCRA. Some of these questions include:

- **Whose culture should be centered and valued in assessment design?** This foundational question arises across scholarly, practical, and policy discussions surrounding SCRA. While some scholars argue that SCRA should create fairer assessments for all learners by expanding inclusivity (e.g., Lee, 2025; Zandvakili & Gordon, 2025), others advocate for explicitly centering the needs and experiences of historically marginalized student groups as a primary goal (e.g., Randall et al., 2022, Randall 2023). This tension becomes more complex as SCRA efforts increasingly seek to account for the nuanced, intersectional identities of learners. For example, although there is broad agreement on the dominance of White, Western/Eurocentric worldviews, there is less clarity about how to represent the experiences of specific subgroups within or adjacent to those dominant experiences—for example, students with particular religious affiliations, those living in rural areas, or individuals whose cultural identities are tied to underrepresented practices, activities, geographies, or communities. Advancing this work will require deeper inquiry into how cultural representation decisions are made within assessment systems—who is involved, what criteria are used, and how trade-offs are weighed. Such efforts are critical not only

for guiding the design and implementation of SCRA-aligned assessments but also for informing broader policy decisions about inclusion, accountability, and equity in education.

- **How do different use cases for large-scale assessment govern appropriate trade-offs for incorporating SCRA into assessment design?** There are many different reasons for including SCRA in assessment design, ranging from humanizing students' experiences with assessment to generating more trustworthy scores. When these reasons are further contextualized by the uses of assessments—both intended and actual—principled decisions about how SCRA is incorporated into different assessment designs can be more effectively made. For example, assessments that are used for advanced placement or admissions decisions may seek to reflect features of sociocultural responsiveness in different ways than assessments that are used primarily to make school- and district-level decisions. The appropriateness of uses might be further differentiated by the relationships among potential users: assessments employed by teachers—who know their students—to review progress with them and their families might attend to SCRA differently than external decision-makers who do not have regular touchpoints with the students being assessed. These distinctions among use cases become even more complex when assessments are used for multiple, sometimes unintended purposes. Nevertheless, making use cases and user relationships explicit can support more intentional, context-sensitive decisions about how to embed SCRA into assessment systems. It can also help identify which aspects of assessment design—such as item content, administration procedures, or reporting formats—require the most focused attention to ensure equitable and meaningful use.
- **How can emerging technologies support SCRA?** As Bennett et al. (2025a) describe, there are significant opportunities for emerging technologies, including artificial intelligence, to better support SCRA. For example, generative AI might support real-time personalization; more efficient generation of item pools that reflect greater linguistic, cultural, topical, and social diversity; the creation of immersive simulations that better capture social reasoning and authentic engagement to support more valid assessments of complex, deeper learning competencies; and the synthesis of evidence across multiple demonstrations of learning to produce more holistic representations of student capabilities. Additionally, AI-powered technologies show promise in

developing flexible scoring mechanisms capable of interpreting responses expressed through varied modalities—such as spoken language, prose, bullet points, graphics, or symbolic representations. These technologies also may help in producing responsive and interactive reporting systems that are better attuned to users' linguistic preferences, implementation settings, and immediate information needs (e.g., using natural language and semantic search to query reports to better understand student performance and next steps; highlighting elements of student performance that are aligned with productive next steps along meaningful learning progressions).

Ultimately, this chapter—and the volume as a whole—underscores that transforming large-scale assessment is both necessary and possible, and is already underway. It is a call to action for assessment developers, policy leaders, educators, and researchers to build systems that reflect a pluralistic vision of learning, one in which all students are seen, heard, and empowered. The future of large-scale assessment must not only measure what students know and can do, but also support who they are and who they aspire to become. If we are to create assessment systems that are truly equitable, valid, and educationally valuable, the work of SCRA cannot be peripheral—it must be central to our reimagining of assessment systems.

References

Badrinarayan, A., & Darling-Hammond, L. (2025). Enacting socioculturally responsive assessment design at scale: Approaches and policies for large-scale systems. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 336–366). Routledge.

Badrinarayan, A., Lyons, S., Miranda, A., & Klyachkina, A. (2025). Leveraging students' cultural and linguistic assets for assessment: A framework for culturally conscious assessment in the Chicago Public Schools. *Educational Assessment, 30*(3), 1–26. https://doi.org/10.1080/10627197.2025.2497775

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment, 28*(2), 88–104. http://dx.doi.org/10.1080/10627197.2023.2202312

Bennett, R. E. (2024). Personalizing assessment: Dream or nightmare? *Educational Measurement: Issues and Practice, 43*(4), 119–125. https://doi.org/10.1111/emip.12652

Bennett, R. E. (2025). A descriptive review of culturally responsive, socioculturally responsive, and related assessment conceptions. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 11–28). Routledge.

Bennett, R. E., Darling-Hammond, L., & Badrinarayan, A. (2025a). *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy*. Routledge.

Bennett, R. E., Darling-Hammond, L., & Badrinarayan, A. (2025b). Socioculturally responsive assessment: Present and future. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 434–442). Routledge.

Ebe, A. (2025). Examining the relationship between the cultural relevance of text and reading proficiency: Using a cultural relevance rubric in reading assessment. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 168–179). Routledge.

Escoffery, D. S., Fletcher, K. E., & Stone-Danahy, R. A. (2025). "A search for my voice": Socioculturally responsive assessment in AP Art and Design. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 262–282). Routledge.

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52*(10), 1115–1124. https://doi.org/10.1037/0003-066X.52.10.1115

Gulliksen, H. (1950). *Theory of mental tests*. Wiley.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kūkea-Shultz, P., & Englert, K. (2025). A path to transforming the assessment landscape. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 306–335). Routledge.

Lee, C. D. (2025). Implications of the science of learning and development (SoLD) for assessments in education. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 29–49). Routledge.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum.

Michel, R., & Shyyan, V. (2025). Accessibility as a core value for locally responsive assessments. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 247–261). Routledge.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report No. RR-03–16). Educational Testing Service.

Mislevy, R. J., Oliveri, M. E., Slomp, D., Crop Eared Wolf, A., & Elliot, N. (2025). An evidentiary-reasoning lens for socioculturally responsive assessment. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 199–242). Routledge.

Moses, T. (2025). Linking responsive assessments: Challenges and possibilities. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 180–198). Routledge.

Nasir, N. S., Lee, C. D., Pea, R. D., & McKinney de Royston, M. (Eds.). (2020). *Handbook of the cultural foundations of learning.* Routledge.

National Academies of Sciences, Engineering, and Medicine. (2018). *How People Learn II: Learners, Contexts, and Cultures.* The National Academies Press. https://doi.org/10.17226/24783.

Nelson-Barber, S., & Trumbull, E. (2025). Assessment for social justice with a focus on American Indigenous students. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 76–98). Routledge.

Raji, M. O., & Baidoo-Anu, D. (2025). Socioculturally responsive post-secondary entrance examination: Implications for equitable assessment design in sub-Saharan Africa. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 399–414). Routledge.

Randall, J. (2023). It Ain't Near 'Bout Fair: Re-envisioning the bias and sensitivity review process from a justice-oriented, antiracist perspective. *Educational Assessment*. Advance online publication. https://doi.org/10.1080/10627197.2023.2223924

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting White supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment, 27*(2), 170–178. https://doi.org/10.1080/10627197.2022.2042682

Sato, E. (2025). Born socioculturally responsive assessment: An approach to design and development. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 115–143). Routledge.

Sireci, S. G., Crespo Cruz, E., Suárez-Álvarez, J., & Rodríguez Matos, G. (2025). Understanding UNDERSTANDardization research. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 415–433). Routledge.

Skerrett, A. S., Pacheco, M., Hinchman, K. A., Cervetti, G. N., Pearson, P. D., & Greenleaf, C. (2025). Socioculturally responsive large-scale assessment: The case of the 2026 NAEP reading framework. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 283–305). Routledge.

Solano-Flores, G., & Ruiz-Primo, M. A. (2025). Cultural validity: Conceptual and methodological considerations for socioculturally responsive assessment. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 50–75). Routledge.

Wang, Z., Sparks, J. R., Walker, M., O'Reilly, T., & Bruce, K. (2025). Group differences across scenario-based reading assessments: Examining the effects of culturally relevant test content. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 369–398). Routledge.

Welch, C., & Dunbar, S. (2025). Designing socioculturally responsive assessments within the mandated bookends of K–12 assessment systems. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 144–167). Routledge.

Zandvakili, E., & Gordon, E. W. (2025). Transforming educational assessment: Implementing equitable and inclusive strategies. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 99–112). Routledge.

Mind Frames for Improving Educational Assessment

John Hattie, Stephen G. Sireci, and Eva L. Baker

Abstract

Assessment in education has long prioritized accountability over meaningful interpretation for learning. This chapter calls for a shift toward assessment in the service of learning, emphasizing insights into student progress, learning strategies, emotions, engagement, and self-regulation rather than just achievement. To support this, educators must develop assessment-capable learners who can interpret and act on assessment results. The authors introduce 10 mind frames to enhance assessment, promoting diagnostic and predictive uses, clear success criteria, instructional alignment, and a classroom culture that embraces errors as learning opportunities. They also explore how technology and AI can make assessments more adaptive and personalized. By embedding assessment within teaching and learning, these mind frames transform it from a compliance tool into a driver of student growth and educational improvement.

This chapter calls for changes in how we think about educational assessments. We believe such changes are needed if assessments will truly serve learners. The changes in how we think about assessments, or as we describe *Mind Frames* for improving assessment emphasize the following *Principles of Assessment in the Service of Learning* (Baker et al., this volume):

1. Assessment transparency provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.

3. Assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition.

4. Assessments model the structure of expectations and desired learning over time.

5. Feedback, adaptation, and other relevant instruction should be linked to assessment experiences.

6. Assessment equity requires fairness in design of tasks and their adaptation to permit their use with respondents of different backgrounds, knowledge, and experiences.

7. Assessment quality and validity should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility.

Moving Beyond Psychometrics Standards

Too often, assessment is undertaken in the name of accountability of schools, teachers, or students–and there is a place for this purpose. However, the title of this volume highlights *assessment in the service of learning.* A central argument of this chapter is this purpose requires a fundamental shift in the ways of thinking about assessment, changing the focus from "Do we have an assessment with the optimal psychometric properties?" to "Are the interpretations from assessments of value, worth, and significance to improving the teachers' or the student's learning?" Yes, these interpretations need dependable measures, but the stopping and starting points are assessment design and the quality and value of the interpretations.

It is fascinating to watch the focus of the *Standards for Educational and Psychological Testing* change over the seven editions since 1952 (see Sireci, 2020, Table 1). The earlier versions assisted psychometricians in developing and defending tests; even in the latest version (AERA et al., 2014), these Standards include almost nothing to defend the users' rights to optimal interpretations. As Hattie (2014) lamented, "The new edition shows only a grudging move from paying the closest attention to the attributes of items and scores toward, at best, a wink and nudge toward standards for the reporting of test information" (p. 34).

If the AERA et al. Standards had been relabeled as the "*Standards for the Interpretation and Uses of Test Results,*" the field would be in a very different and happier place. If the focus were on interpretations and uses, then teachers and users would more likely attend to the *Standards*, there would be much more discussion about the validity of interpretations and uses, and the field of score reporting would be deep, long, and much further advanced than it is now. Although we are pleased to see a chapter on score reporting finally made it into the soon-to-be-published edition of Educational Measurement (Zenisky, O'Donnell, & Hambleton, in press), we believe 'Interpretation Standards' would assist tests in being seen as worthwhile to users and not as a set of rules for test developers alone. We also believe interpretation standards are supported by the transparent design of assessments to support instruction.

This shift, of course, is not denigrating the importance of creating and validating tests, but if tests STARTED from the nature and depth of valid reporting we wished to make, the liking and development of our profession would be more effective in using assessments to serve learners. Most teachers would know and welcome the *Standards*, most policy makers would not continue to ignore the fundamental premises of the *Standards*, and 'report design' mavericks that flood the internet with glossy reports would have a basis for defending their claims.

Technology has advanced such that so many companies promote glossy reports promising massive improvements in student performance, and it is rare now to expect a user to read a manual or note cautions about the need to triangulate findings. We have seen so many 'reports' placed in folders never to be acted upon; there are so many adverse reactions to 'testing' (e.g., 'we teach students, not data'), and so much data collected that are not used—but schools continue to collect data

because they "have to" and the resulting data and scores provide comfort that "evidence is available, upon call."

We need to do better, and we can. Doing so will require adhering to the standards we set for ourselves, prioritizing using assessment to promote learning over accountability and adopting several other "mind frames for assessment," which are the focus of this chapter.

Shifting the Focus to Improvement

There is value in accountability models of assessment in informing policy makers of the impact of curriculum reforms, investment of resources, and needs for the future. However, classroom assessment's main functions—to inform and improve teaching and learning—are more important. The claim is that assessments can be crucial aids to inform and improve teaching and learning, help educators see "their impact," and accelerate learning when used for diagnostic and predictive purposes.

There needs to be interpretation relating to diagnosis (Where are the students now?), progress (How are they going?), informing next steps (Where should they go next?), and summatively, to ascertain who has attained the success criteria of the lesson(s). Classroom assessment can thus be considered descriptive, ascriptive, and prescriptive. Primarily, the power of assessments comes from feedback to teachers about their impact: What did they teach well, and what not? Who did they teach well, and who they did not? And were the gains made appropriate for the time and resources invested? That is, what, who, and how much. Then, students are the greatest beneficiaries when these three impact questions form a primary assessment function.

Being clear about the diagnosis (including prior performance and current status) and desired success maximizes the power of assessment to know how to accelerate growth best. There is rarely one right way between these two points, and here is where the art of teaching is so critical, especially to evaluate continually the impact of the choice of teaching, the concepts and misconceptions, learning from the errors that are made (they need to be teaching opportunities, not embarrassments, Hattie, 2023). Also, differentiation fundamentally means different ways and times to attain success. Optimal instruction relies on understanding where a student is on a learning progression, the ability to probe for understanding and misunderstandings, and the provision of guidance toward the next steps.

Success involves more than knowing the content, but also includes logical reasoning, understanding cause and effect, organizing prior and new knowledge, problem-solving, and mastering subject matter knowledge and principles. Success involves the 'knowing that" (content and facts), the 'knowing how' (the patterns and deeper conceptual understanding), and 'knowing with' (transfer to near and far situations) (Hattie, 2023; Roland, 1968; Ryle, 1945).

This refocusing of assessment places much emphasis on reliable and dependable interpretations, and these need to be open to critique, moderation, and triangulating (from the teacher noticing, from students' voices about their learning, from others' observing the teacher's impact, and from other measures). It is noted, however, that 'assessment' occurrences can be events (test or assignment) or in-passing, and these need to be planned to inform the teaching status and progress.

This shift in focus aims to develop student assessment capabilities to drive their own learning (Fisher et al., 2023). That is, to teach students how to interpret the results of any assessment such that they can (a) learn to be involved in making decisions about Where to next?, (b) understand more what they know and can do and what they still need to learn to attain success; and (c) fundamentally hear, understand and act upon the interpretations their teachers and they make to improve the rate and depth of learning. The current debates about whether assessments should be graded or marked seem trivial—as the received and understood interpretations matter and not what form they are delivered.

In the remainder of this chapter, we propose 10 *mind frames* to inform thinking about assessment so the power of diagnosis and interpretation can be realized, and assessment can truly be used to accelerate learning.

The Ten Mind Frames for Assessment

Mind frames are ways of thinking related to the cognitive patterns, perspectives, interpretations, evaluations, and mental models we use to interpret our world. They influence how we perceive information, make decisions, solve problems, and navigate with others. What we do matters less than how we think about what we do. In schools, for example, this thinking is the precursor to choosing high-impact strategies, ensuring the fidelity of implementation, and evaluating if there have been important impacts on students from the teacher's delivery of this instruction. It is the teacher's thinking that leads to the choice of interventions, devising and

explaining the learning intentions and success criteria, knowing when a student is successful in attaining those intentions or not, having sufficient understanding of the students' understanding that they bring to the task, and sufficient knowledge about the content to provide meaningful and challenging experiences in various progressive pathways to success in learning.

Mind Frame #1: Assessment in schools needs to consider both progress and achievement, learning emotions, and strategies.

Psychometrics has focused on modeling tests that rank-order students and have little connection to the classroom. We can trace the extraordinary history of item response theory (IRT) back to Thurstone (1925) and Lord's (1952) doctoral dissertations (Luecht & Hambleton, 2021; Thissen & Steinberg, 2020). Most of these models and applications were based on ability testing, and there has been (more in the early days than recently) a robust discussion of whether the assumptions can ever be met when using achievement data (Traub & Wolfe, 1981). There are many problems in using IRT (an essentially norm-referenced model) for achievement testing. Often, there is a major restriction in the range of student proficiency as measured by the test, the assumption of unidimensionality is unlikely to be met (e.g., an achievement test might exhibit unidimensionality when the problems are relatively novel for students, but not as they become more expert; Snow & Lowman, 1984), and there is more often a non-normal distribution of achievement. Moreover, teachers rarely calibrate items, check for even minimal psychometric properties, and most could not even spell IRT and know little about classical test theory. They attempt to compensate for these losses with frequent testing, triangulation with observations, and assessments used to motivate or grade students. In addition, there is a loss of confidence in using many school assessments, especially for reasons of accountability. It is hardly surprising that the major advances in measurement over this past century have rarely crossed the school gate.

When we consider "tests" in schools, the dominant focus is achievement, and all too rarely are there measures of progress, emotions, learning, climate, striving, or engagement. The fundamental claim of this first all-important mind frame is that it is via a high trust, inviting, and safe climate that students are prepared to be challenged to know that which they do not already, engage in error management as opportunities to learn (not embarrassments), be taught to choose optimal learning strategies for the tasks such that there is progress towards higher achievement

relative to their starting position. Instead, the prime focus on achievement often identifies the high and low achievers, and then explanations are sought about why some students can and some cannot.

A fundamental thesis is that educators need to start with high-trust climates, teach learning strategies, and then focus on their students' progress to higher achievement. We should not start with achievement, as this distorts the conversations to privilege those who begin well above average, and enhanced achievement (no matter where the student starts) is supposed to be an outcome of schooling.

The emphasis on high achievement leads to perverse consequences for policy makers, parents, educators, and students. Too many believe a "high achieving school" is necessarily a "great school." If the students start above average, many of these high achieving schools can support 'cruising'—that is, adding no value (e.g., across Australian schools, almost 50% are in this cruising mode; Hattie, 2019). If students start below average and gain more than a year's growth for a year's input, this is stunning and needs esteeming (even if the final achievement is still not above the state or country average). Beliefs in 'high achievement' beliefs damn those who start below average and make substantial progress. We need both high achievement and high progress (See Figure 1). We get high achievement from high progress, and a simple achievement by progress chart could transform the way we make interpretations about what is optimal, how to advance learning and learners, and show that those teachers who impact progress are much more expert and should be esteemed than those who defend cruising. The aim of schooling is to continually move all students from left to right (and by arithmetic, more will increase from lower to higher achievement: every student, no matter where they start, deserves at least a year's progress for a year's input.

Any scores from classroom assessments administered over time can be plotted (and effect size and other measures of progress determined); students thus become their own baselines, and all can see who made progress or not. The interventions to improve will likely differ depending on which quadrant the student is located. Consider two high-achieving students, one in the Cruising and one in the Optimal quadrants. Conventionally, all would be happy that they scored high—but one of these students is cruising, and the other is improving their learning—hence, the teaching strategies need to be different. If only achievement is considered, both seem to be succeeding in the class—but clearly, this is not so for the cruiser.

Similarly, for the two low-achievement students, one is progressing and one is not, and the student progressing should be esteemed similarly to the above-average student who also is progressing. The progression, more than the achievement, is critical for successful schools and learning. Of course, when high progression and high achievement are the outcomes, this is also a highly desirable state.

Measures of learning

When we ask teachers about their 'theory of teaching' this is a prolonged, profound, and plentiful conversation. When we ask about 'their theory of learning' it is too often short, shallow, and subjective. This gap underscores the importance of providing such measures of learning strategies to teachers not to classify them into styles, groups, or hierarchies; but to learn which strategies they use, whether they have fallback strategies when their first choices do not work, and whether their students have the self-regulation skills to choose the optimal learning strategies aligned with the requirements of the task. These are all difficult skills, but essential for successful learning for the specific tasks. There are some learning strategy scales (see Schellings, 2011), but what students self-report may be more what they think they do, not what they do; and may reflect their wishes and beliefs about how they learn. Thus, more sophisticated measures may be needed. Even as adult educators, we struggle with a language about how we learn. Hence, what chance do students have trying to discover these secrets? More powerful measures may include 'thinking aloud' measures, biometric analyses, and it is hoped that the developments in the science of learning can inform the best forms of measuring these skills (Hattie et al., 2024). Other attributes worth developing include measures of cognitive load, retention and forgetting measures, and measures relating to the skills of not only acquiring knowledge and understanding, but skills of consolidating these into chunks or using pattern recognition to better retain for transferring to near or far contexts. All of these efforts need to be grounded in the design of assessments integrally linked to learning goals and processes.

With the push by employers for graduates who can work in groups, translate their knowledge to others, and lead and teach others, schools that do not develop these group skills are not assisting their graduates in being employed (Deming, 2017). There are fascinating measurement problems in measuring an individual's contribution to the group along with group scores and identifying and measuring the 'I' and 'We" skills necessary for successful group functioning (Hattie et al., 2021).

Measures of learning emotions

Students experience many emotions in their experiences in classrooms. Positive emotions, such as happiness, curiosity, and excitement, can enhance learning by increasing motivation, attention, and memory. The most negative influences relate to student anger, procrastination, depression, and anxiety; but the dominant negative emotion is boredom (Blannin et al., 2025; Moeller, Brackett, Ivcevic, & White, 2020). Positive emotions can increase motivation to learn and willingness to take on new challenges, while negative emotions can decrease motivation and lead to lower motivation to engage in learning activities.

Measures of motivation

Many formal and informal measures of motivation involve the observation of students. Early researchers emphasized types of motivation, such as extrinsic and intrinsic rewards. Extrinsic rewards have many limits, including the important idea that the learner must find personal value in the reward. Moreover, there are a range of extrinsic consequences for learning, some short-term and others more delayed. If, for example, good grades in school do not matter to students, they obviously will not inspire improved performance. Intrinsic rewards are thought to engage students by encouraging them to internalize the value of what they are learning. They learn a domain for their own satisfaction. Consider the student who loves history or biology and wants to learn it for its own value. There is an extensive literature on this topic (see Ryan & Deci, 2000) for an excellent review).

More interesting approaches have addressed a specific outcome connected to both archetypal types of motivation. One of these involves the consequence of self-efficacy, whether learning increases the student's capacity to learn or perform (Bandura, 1993). Bandura identified four pillars of self-efficacy development, including 1) mastery experience or performance outcomes (taking on a challenge and being successful;) 2) vicarious experiences (social role models) where the learners see others like themselves succeed: 3) social persuasion, involves positive feedback during learning; and 4) emotional and physiological states; to which we add a fifth, evidence of learning begets more engagement towards more learning. Bandura (1986) noted that being in an environment of wellness may support affective arousal and willingness to learn. It is also clear that a cultural component influences motivation, including the concept of locus of control, the belief that one can change performance by effort compared with the idea that outcomes are predetermined by level of “intelligence” or luck (see Sagone & De Caroli, 2014).

How does one know if a student is "motivated"? Observation has limits, particularly in a teacher or researcher's ability to draw trustworthy inferences from disparate student behaviors. Just because a student is seen to be 'doing' the work is no guarantee of successful learning progress. Short-form self-report measures have been used to determine students' responses to instruction or topics, their willingness to voluntarily seek more of what they have been learning, or their willingness to recommend the topics to a friend. These brief, teacher-made tests, best administered anonymously, can give rapid information about likes and dislikes, interests and self-efficacy regarding lessons. In computer-supported learning, including games, sensors associated with arousal and engagement involve eye-tracking devices, inferences drawn from delays, and other more sophisticated algorithms. Plass and Kalyuga (2019) present a contemporary summary of these efforts. The onset of more artificial intelligence options should allow teachers to make better inferences from students' behaviors, although appropriate applications are on the horizon.

Summary of mind frame #1

This first (and most critical) mindset aims to best inform educators about their impact on students' learning experiences so they can make better diagnostic interpretations leading to more effective instruction and experiences to accelerate students towards the success criteria. The message is to not only look at achievement, but also the causes and correlates of achievement, such as the high trust, high expectations and inviting climate, the choice of learning strategies, the emotions that speed or impede the experience of learning, the consequential levels of engagement, skills in working alone or in groups, and the motivations to want to learn more and deeper about the topics we desire them to learn about.

Mind Frame #2: Develop students' assessment capabilities so they can interpret the feedback and 'where to next' from assessments.

This mind frame emphasizes that we need to ensure students can (a) interpret the assessment results correctly and (b) make some consequential actions, decisions, or thinking that informs their next steps in learning. Sadler (1989, p. 143) noted "it is insufficient for students to rely upon evaluative judgments made by the teacher"; thus, requiring students' critical engagement in discerning the quality of their work and the criteria and standards against which their work is being judged (see also Baird, 2014).

Absolum et al. (2014) recommended all students be educated in ways that develop their capabilities to assess their own learning, and that the success of any national assessment strategy be judged by whether all students are developing the capability and motivation to evaluate, interpret, and use information from quality assessments in ways that affirm or further their own learning. Too often, it is the adults who make the interpretative decisions, and this is as it should be when it informs the impact of and the next decisions by teachers as to what next to teach. However, they claimed that teaching students how to interpret the results of any assessments is necessary to enhance student learning. This permits them to access, interpret, and use information from quality assessments in ways that affirm or further their learning.

Frey et al. (2019) argued that developing assessment capable learners leads to students becoming more aware of their current level of understanding in a learning area, more keen to understand their learning path and have renewed confidence to take on the challenge, better at selecting tools and resources to guide their learning, more ready to seek feedback and recognize that errors are opportunities to learn, more able to monitor their own progress and adjust course as needed, and recognize what they are learning and can teach others (see also Fisher et al., 2015). Thus, making students assessment-capable is perhaps one of the best ways to transform assessments into tools of learning.

To enable students to make these interpretations, teachers must model ways of using assessment information that helps students to meet their learning goals. In this way, students learn about: setting and clarifying challenging learning goals; how to access, interpret, and use evidence; understand the dimensions of engagement that lead to better outcomes; and engage in evaluative thinking about whether the work is good enough, meets the success criteria, and where to make the next learning moves. Wyatt-Smith and Adie (2021) suggested engaging students in discussions and activities to reach a shared understanding of the purpose of the assessment, learning goals, and judgments, applying this understanding to feedback to improve work and develop learning goals, engaging in self- and peer assessment and feedback, co-constructing or deconstructing criteria with teachers, peers or self, build content knowledge and skills to enable decision-making about the quality of one's work, and engage in dialogue with teachers regarding the student's areas of strength/weakness, and learning goals.

These recommendations correspond with the vision outlined by Gordon and Rajagopalan (2016). Specifically, when assessment focuses on improving teaching and learning rather than measuring only what students have learned, we can achieve excellence in education for all students in America. They argued that the challenge for the American education system is not to determine whether or even by how much students have failed to achieve, but to enable them to learn and develop as fully as they are able so they can navigate the world around them, live fuller lives, and contribute as fully as they can to society.

Summary of Mind Frame #2

Students can be empowered and have greater agency in their learning if they understand assessment results and how they lead to next actions. Teachers can help students acquire this understanding by modeling assessment information can help students meet their learning goals

Mind Frame #3: Formative, summative, and ascriptive evaluation refers to the timing and nature of interpretations (and NOT anything to do with kinds of assessment).

Scriven (1967) introduced the concepts of formative and summative evaluation, which are not intrinsically different types of evaluation but have different purposes. Formative evaluation is designed, done, and delivered to make improvements to the evaluand, and summative evaluation is done for, or by any decision-makers who need evaluative conclusions for any reason *other than* conceptual development. The key concept is that in "light of the (formative) processes, or some of them, the product is (or is not) finally revised and released, and summative evaluation begins" (Scriven, 1993, p. 3). So, the distinction is that the purpose is formative during and summative at a key milestone: when the cook tastes the soup it is formative, when the guests taste the soup it is summative (Stake, cited in Miller et al., 2016). Both are important, both need to be appropriately rigorous, neither is more worthwhile than the others, both depend on the quality of interpretations, both require judgments.

One of the major fallacies is that there are concepts such as formative and summative **testing**. Bloom et al. (1971) applied Scriven's terms to education and learning with the release of their book *Handbook on Formative and Summative Evaluation of Student Learning*, where the terms were intertwined with assessment. Give us any test, and we can make formative or summative interpretations. A test

is neither formative nor summative; it depends on when an interpretation is made for improvement or the end of an intervention. It may be that tests that lead to more optimal formative interpretations more closely track instruction and a level of detail to allow sub-tasks to be addressed. In contrast, summative evaluations are more about whether the intentions of the lesson are known and understood. Still, both require defensive psychometric properties, interpretations (to improve or to ascertain status), and the difference is more in the timing and purpose.

This yoking or formative and summative to 'tests' has led to a focus of many professional development programs. For example, Black and Wiliam (2010) noted "formative assessment is an essential component of classroom work" and that "they know of no other way of raising standards for which such a strong prima facie case can be made" (p. 14). Perhaps it is unsurprising that Wiliam later argued, "the biggest mistake that Paul and I made was calling this stuff 'assessment'… because when you use the word assessment, people think about tests and exams" (Booth, 2017, p. 2). He argued that the program may have been more accepted and successful if they had used "Responsive teaching" and not tied it to 'tests.'

It is time to revert to Scriven's original claim: formative and summative evaluation–refers to evaluative thinking and evaluative decision-making when reviewing the outcomes of teaching, learning, and assessment opportunities (Clinton & Hattie, 2024). The quality and defense of the interpretations become critical (and not the test, sui generis). The tie-in with assessment has done a major disservice to the original Scriven notions. It has led to too much emphasis on the assessments and too little on the timing and quality of interpretative information.

Summary of Mind Frame #3

We need to abandon notions of formative and summative tests because those terms are misnomers. Educational assessments can give us formative and summative information, and that it is the timing and nature of the interpretations that make them valuable.

Mind Frame #4: There are at least three levels of knowing: Knowing that (concepts, ideas, facts, surface), knowing how (relations, deeper conceptual), and knowing with (transfer, pattern recognition)—and assessment may need to measure each separately.

Contrary to many test developers' claims, a majority of the items on standardized assessments in the USA can be answered by simply knowing lots, especially in knowledge-rich subjects like science and history (Koretz, 2017). In fact-dominant classrooms, teachers ask questions primarily about the facts, and students soon realize that 'knowing lots' is the sign of a good learner in many classes. However, most theories of schooling ask for more than knowledge-rich graduates. There is also a clamor for students who can see patterns across ideas, have deeper conceptual knowing, and can transfer ideas from one to another situation or problem. The need is not either facts or deep, but facts *and* deep—depending on the nature of the problem. Much deeper and readily available analyses are needed as to how users actually answer items, the knowledge and pattern recognition they use, and the error management they use to solve a problem and decide on their optimal answer.

One of the greatest travesties has been an over-reliance on Bloom's (1956) taxonomy. It mixes knowing (knowledge, comprehension), ways of knowing (analysis and synthesis), outcomes (applications), and evaluating and creating. There is no hierarchy; everything fits somewhere (which means everything is ok), and there is limited to almost no research on the value of the taxonomy (Hattie & Purdie, 1998). In 2001, Anderson and Krathwohls (2001) revised edition was published, acknowledging many of these deficiencies and adding another more powerful dimension: learning objectives for each of the six' levels': factual (what needs to be known, conceptual (interrelationships among the basic elements); procedural (how to do something, methods of inquiry, and criteria for using skills, algorithms, techniques, and methods), and metacognitive (awareness and knowledge of one's cognition). These latter three can be termed degrees of cognitive complexity, and they have been developed using many learning and knowledge models.

Webb (1997, 2002, 2007) developed the 'Depth of Knowledge' model, and it has four levels quite similar to the revised Bloom new dimensions: recall, skill or concept, strategic thinking, and extended thinking (also see Hess, 2006; Francis, 2022). Recall relates to reproducing a fact, principle, or routine procedure. Skills or concepts relate to using information, selecting appropriate procedures for a task, or organizing and displaying interpreted information. Strategic thinking involves reasoning or developing plans to approach a problem, employing decision-making and justification, and solving abstract, complex, or non-routine problems. Extended thinking involves performing investigations or applying concepts and skills to the real world that require time to research, problem-solve, and process multiple conditions of the problem or task. Webb's Depth of Knowledge model relates to the depth of content understanding, the scope of a learning activity, and the skills required to complete tasks.

A powerful taxonomy is SOLO (Structure of Observed Learning Outcome) taxonomy. Developed by Biggs and Collis (1982), it has been critical in developing assessments, scoring, rubrics, and teaching and learning. The SOLO taxonomy consists of five levels of increasing complexity: pre-structural, unistructural, multistructural, relational, and extended abstract (simply referred to as no idea, one idea, many ideas, relating ideas, and extending ideas). Each level represents a different level of understanding and the ability to think about and use information in increasingly sophisticated ways. Thus, it is possible to conceive of 'difficulty' as an increasing challenge within each of the four levels, and 'complexity' as an increasing challenge when moving from instructional to extended abstract.

The SOLO taxonomy provides a valuable framework for understanding and assessing learning outcomes as it is based on a model of cognitive complexity that allows educators to identify the level of understanding that students have reached and to design appropriate learning activities and assessments that challenge students to progress to higher levels of understanding (know where a student is performing and aim one step higher). Students can also use it to reflect on their learning and identify areas where they need to improve their understanding.

Clinton et al. (2021) reviewed Bloom, Depth of Knowledge, and SOLO and developed a model that brought all three together. Their four levels are categorized into two attributes: 'knowing that'/surface and 'knowing how'/deep thinking. The first major attribute of cognitive complexity relates to *knowing that* or surface thinking—the ideas, the factual, and content knowledge. This includes:

1. Factual knowledge recall and reproduction: for example, 'I know and can distinguish various parts of human hand anatomy.'
2. Conceptual knowledge involving basic application and skill: for example, 'I am able to apply Van Gogh's painting techniques to my drawing of the Sydney Opera House.'

The second major attribute is *knowing how* to think deeply or develop relations between ideas, extending to near or far new situations.

3. Strategic or relational thinking: For example, 'I understand that empowerment evaluation design principles have been utilized to design this evaluation compared to other evaluation models.'
4. Transfer: for example, 'I will be able to apply and adapt my methodological learning from my previous understanding of jazz principles to this new piece of music.' This can also be termed *knowing with.*

A summary of the three cognitive taxonomies using Clinton's (2021) model is presented in Table 1.

This powerful way to distinguish the levels of cognitive complexity was suggested by Ryle (1945), who distinguished between 'knowing that' and 'knowing how.' 'Knowing that' is the knowledge that something is the case (e.g., knowing an evaluation theory), and 'knowing how' is the knowledge you have when you know how to do something, such as how to ride a bike, bake a cake, or make an evaluation interpretation—"how to make and appreciate jokes, to talk grammatically, to play chess, to fish, or to argue" (Ryle, 1949, p. 28). 'Knowing how' cannot be defined in terms of 'knowing that,' nor is 'knowing how' necessarily logically before 'knowing that' (Kapur, 2008, 2012, 2016; Oberauer, 2010). For example, you cannot teach a novice chess player how to play to the same standard as the grandmaster just by feeding the novice facts about the game. Thus, 'knowing that' entails 'knowing how' to put the 'knowing that' into practice, but 'knowing how' cannot be built up from pieces of 'knowledge that.'

Clinton et al. (2021) argued that underlying these four levels is the notion of "evaluative thinking" or self-regulation–knowing when to be surface and when to be deep, and this skill invokes a particular kind of critical thinking and problem-solving. Evaluative thinking is the process by which one marshals evaluative data and evidence to construct arguments that allow one to arrive at contextualized value judgments in a transparent fashion (see also Buckley et al., 2015; Lee, Wallace, & Alkin, 2007; Vo et al., 2018).

Summary of mind frame #4

Assessments need to be developed, scored, and reported at these four levels, or at least at the "Knowing that" compared to the "Knowing how and with" levels. This is not to be confused with difficulty at each level (factual, strategic, etc.). There can be increasing difficulty levels, but cognitive complexity moves down the depths from surface to deep to transfer. These levels were the basis for developing the NZ elementary and high school e-asTTle (https://e-asttle.tki.org.nz/; Hattie et al., 2006), where assessments were automatically created using linear programming methods (van der Linden, 2005) so teachers and students could understand how they performed on both easy to more difficult items and on surface to deeper cognitive complexity.

Mind Frame # 5: We care about alignment between standards, success criteria, lessons, tasks, and assessments, and recognize the power of backward design (e.g., reports to tests, summative goals to influence formative directions, etc.).

A fundamental assumption of unidimensionality is that the attribute being measured can be ordered from high to low, more to less, effective to ineffective, etc. But the world of classrooms is multidisciplinary with various teaching methods, learning strategies, safety to learn and err, curriculum progressions, tasks, and assignments –oftentimes, the outcome is also multidisciplinary. This often leads to simplistic dichotomies or decisions that simplify but distort the complex nature of reality. It is not necessarily either-or, but when, how, and with what impact. Surma et al. (2024) make a strong case for both rather than either-or claims about typical dichotomies such as knowledge-rich vs. deeper learning, direct instruction vs inquiry teaching, multiple choice- vs. open-ended, grading vs. comments.

A fundamental mission of education is to influence academic outcomes (alongside many other attributes noted in Mind Frame 1). When trying to understand the underlying reasons for the very discrepant effect sizes of many teaching methods, Hattie (2023) developed the 'Intentional Alignment model'. This model relates to the proportion of knowing that, how, and with (surface, deep, transfer) in the lessons, and then aligning these the notions of success (e.g., one success criteria for knowing that, and one for knowing how and with), the methods of teaching, the optimal strategies of learning, the cognitive requirements in the activities, and the assessment methods.

There are seven major parts of the Intentional alignment model:

1. Determining the learning intentions and success criteria, typically based on a curriculum, understanding progressions (where the students have been progressing, where they are now, and where they need to be), and the motivations and dispositions students bring to the class.
2. Cognitive task analysis of the knowing that, knowing how, and knowing with foci of the lessons (i.e., content, deeper understanding, and transfer skills and knowing). Understanding the cognitive complexity involved in learning ensures students have the strategies for learning, understand their progress, and have a concept of what success looks like.
3. Creating a climate and culture of the class to ensure students and teachers see errors or misconceptions as opportunities to learn. There needs to be safety in working, discovering, and exploring with peers, an inviting climate for all, and acknowledging the diversity of what each and all students bring to the class environment.
4. Teaching methods that align with the cognitive complexity of the various components of the tasks, the required confidence to take on the challenges of this complexity, and efficiently and effectively improve students' progress towards the success criteria.
5. Ensuring students have appropriate and effective learning strategies to engage in the complexity of the tasks.

6. Choosing activities that align with the content's complexity levels, the deeper conceptual and relational thinking.
7. Using assessment methods that focus on the content ('knowing that'), the relational ('knowing how') and transfer ('knowing with') that inform teachers and students of their progress, success, and gaps to be then addressed.

These aspects of alignment are not only critical for using assessment to promote learning, but they are also essential for providing (a) evidence of the validity of these classroom assessments, and (b) professional development for teachers as they improve their teaching and formative evaluation processes. As the AERA et al. (2014) *Standards* stated, "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). The curriculum and goals of instruction provide the theory to support test score interpretation and use, and confirmation of alignment provides the evidence to support such interpretation and use. Moreover, teachers develop and strengthen valuable skills for helping students learn by developing and sequencing assessments aligned to instructional goals.

It would be worthwhile to at least have separate tasks, assignments, and assessments for the 'knowing that' and for the 'knowing how/with.' This would make transparent to students what is being valued, make the feedback from any assessments more easily understood, and encourage the use of differentiated teaching, learning, and activities depending on the focus on knowing that or how/with. Hence, the validity of interpretations from assessments is more robust.

Summary of Mind Frame #5

Alignment is needed across all stages of the education process, from goal and standard formulation to assessment, interpretation, and feedback. The Intentional alignment model provides a comprehensive way for building and evaluating such alignment.

Mind Frame #6: Provide scoring rubrics and success criteria near the beginning and not the end of lessons/courses to provide verisimilitude to any assessment.

Imagine asking students to play their video games and NOT telling them what it means to be successful or what it means to get to the next level. They would say it is pointless, so why would I engage in activities without a purpose or goal? But this is often the case in schools when students are told to "do the work" with little to no understanding of what is 'good enough,' what the criteria of success look like, or now know the evaluative decisions to be made to claim success. Blannin et al. (2025) analyzed the results from an App that asks students about their experience in class, and the dominant emotion is boredom. However, when students can explain what they are learning, there is a 73% (from 37% to 64%) improvement in enjoyment and engagement in their learning.

There must be appropriately challenging expectations embedded in the success criteria. That is, challenging relative to what each student cannot do (yet), and then the teaching focuses on enabling students to engage in the challenge. The Goldilocks principle of challenge is optimal: not too hard, not too easy, not too boring. Unfortunately, some students are reluctant to take on challenges (for fear of failure, becoming an embarrassment in front of peers and the teacher, or not having the skills to address the challenge). Hence, it is important to specifically know, plan for, and develop confidence to take on appropriate challenges (self-efficacy), and have high levels of trust that failure is the best friend of learning in this class.

As noted earlier, most lessons may need two success criteria: one for the 'knowing that' and one for knowing how and with' goals): similarly with assessments. It is a balance of proportions, not an either/or. The effectiveness of assessment can increase when students are provided with scoring rubrics. Scoring rubrics are more effective when they are provided at the outset rather than at the end of activities, and even better when students are involved in creating them (Becker, 2016), when they relate to lessons over time (and not short term(Andrade & Valtcheva, 2009).

Summary of mind frame #6

Teachers know what is coming next in an instructional sequence and how they will evaluate student work. Students should know these things also; so they and their teachers have the same target in sight. By providing students with the rubrics with which their work will be evaluated, they will approach the tasks more confidently, and thus enhancing their learning.

Mind Frame #7: We create a climate of welcoming errors and not knowing; thus, a major role of assessment is diagnosis and discovering where students are currently knowing and not understanding.

Mastery is often claimed to be 80%+; when a student gets 100%, perhaps the test was too easy; when they get <40% they are usually despondent. Testing would have little use if it did not detect what the student does not know and understand, what the focus of the next phase of teaching and learning should be, and if the climate is such that there is negativity from learning from errors and not knowing (Law et al., 2004). Indeed, we do not come to class to learn what we know and understand, so the starting point is 'not perfection.' Sadly, so many learn quickly that classrooms privilege those who know and can do, eagerly respond to teacher questions, and know how to do the work. A major role of assessments needs to be diagnostic and discovery of where students are currently, and then make predictive recommendations for further learning. Too much assessment is making a status report on where 'they are now,' which is valuable, but the starting point to 'where to next' is the essence of learning.

Feedback thrives on errors, and undertaking challenges leads to errors. Therefore, the climate of the class must welcome errors, teach students how to engage in error detection, have teachers positively engage in error repair, and failure needs to be a learner's best friend. Errors can serve as important feedback information, indicating where the student's thinking and knowledge are not yet developed. However, students and teachers often shun errors to avoid negatively impacting a student's self-esteem; peers can be nasty, brutish, and short to those showing they do not know. When students make errors in classroom discussions, they are usually quickly corrected (by the teacher or by the teacher asking a peer), and many students soon learn that it is better to look like they know and hope they do not get asked. Thus, avoiding and withdrawing are successful tactics to maintain their sense of themselves as learners.

Many learning models explicitly include attention to errors and failure. For example, Piaget's (1952) discussed cognitive disequilibrium, which occurs when learners encounter a situation contrary to their current mental model (such misconceptions have been studied mainly in science learning). Then students are challenged or instructed until they either assimilate those differences into their mental model or modify it according to the new information. Productive failure invites students to solve, usually ill-structured or complex, problems before instruction to evaluate

their disequilibrium, try and invent multiple solutions, and realize what they need to learn from the subsequent instruction (Kapur, 2024). Thus, there needs to be a climate for receiving, welcoming, and learning from test information.

Gordon (2020) has long promoted assessment in the service of learning. This involves interrelating assessment, teaching, and learning such that they are reciprocally employed each in the service of the other. Measurement is no longer primarily for testing but to inform "teaching and learning transactions, their outcomes and their continuing assessment" and thus can a) measure the status of developed ability; analyze processes of teaching and learning; understand intentions, appreciations, needs meanings, performances; and cultivate learning and development of abilities, appreciations, competences and skills (p. 73). Assessment can not only measure but also cultivate learning.

Summary of mind frame #7

Classrooms should embrace a culture where mistakes are welcomed and discussed. In such environments, the diagnostic role of assessments can help students discover where they are and what to do next improve their learning

Mind Frame #8: We act on the belief that a major purpose of assessment in schools is to inform teachers of their impact.

This is the major message of the Visible Learning work—we care less about how teachers teach and assess but care more about the impact of their teaching and assessment interpretations. Switching from the act of teaching and testing to the *impact* changes the debate, makes it easier to recognize expertise, and puts the focus on the dependability of the interpretations and consequential actions. Excellence, for example, is not tied to the use of any specific set of teaching strategies or testing methods but to the optimal alignment and fidelity of implementation of strategies and methods that have an impact on student learning. This focus on impact means interpretation of multiple sources of evidence, including test scores, is critical. Other sources include teacher-set assignments and assessments, their noticing, student voice about their learning, the evidence of the safety to not know and make errors, and classroom management to include *all* in the learning. The key is triangulating this evidence from tests, teacher noticing, student voices about their learning, and artifacts of student work. Every school has pockets of high-impact teachers with high levels of evaluative thinking. The

core question is how to understand this evaluative thinking such that it can scale up. In education, we are good at finding and fixing problems, but not so good at identifying excellence and scaling it up (Baker, 2004).

Summary of mind frame #8

Educators can use assessment to assess their impact and the impact of the different teaching tools and instructional practices they use.

Mind Frame #9: Many technologies (especially large language models) can make developing and scoring assessments more efficient and lead to making assessment interpretations more defensible.

The most remarkable change in our lives is the advent of large language models such as ChatGPT, Gemini, Claude, etc. These tools offer the most remarkable transformation of assessment since the development of IRT. While the usual claims about evaluating the dependability and validity remain (and become perhaps even more important), these AI models can more readily 'write' items to specifications, offer more opportunities to adapt tests on the fly, score open-ended assessments, and create more tailored interpretations from the assessments. It is early days, but the possibilities abound. Maybe the over-dependence on multiple-choice and closed items that have dominated many assessment systems (as they are cheaper and easier to score) will be reduced, and items that map the processes of learning, the deeper ideas, and the construction (rather than recognition) of answers will take their rightful place within tests.

As students gain more insight into how they can create, score, and interpret assessments via AI, it will be important to understand the skills we need to teach them to do this in a worthwhile way to enhance their learning. Such skills could include how to ask probative questions (if the wrong prompts are asked AI still gives answers), assessment credibility (are the AI comments and recommendations right or wrong), evaluative thinking (are they 'good enough'), making wise choices (these tools can make recommendations for next learning steps and they may or may not be optimal), oral fluency (you can speak to the AI engines), and collaborative critique (how to engage others in critiquing the use and outputs from these tools).

There have been significant changes in assessment over the last decades due to technology. These changes include automated scoring of constructed responses, digital assessments that provide log (process) data that can be used to understand

better test takers' cognitive processes, advances in computer adaptive testing, personalized or designed-in-real-time methods of administration, and improved score reporting that maximizes the interpretations from the assessments.

Sireci et al. (2024) proposed a new model using many of the advances from AI tools called Design-In-Real-Time (DIRTy) assessment, which reflects the progressive evolution in testing from a single test to an adaptive test, to an adaptive assessment system. It involves: (a) assessment building blocks (individual items or "assessment task modules" (ATMs) that are linked to multiple content standards and skill domains), (b) gathering information on test takers' characteristics and preferences and using this information to improve their testing experience, and (c) selecting, modifying, and compiling items or ATMs to create a personalized test that best meets the needs of the testing purpose and the individual test taker.

What is new in DIRTy assessment is tailoring the test to multiple, personal factors, and delaying test specifications until the interaction of a test taker and a testing purpose occurs. DIRTy assessment, can use individual items or sets of items aligned with both content standards and job tasks that represent the building blocks of a unique instantiation of an assessment, and a system to search and assemble those sets of items (ATMs) to meet the additional configuration goals of personalized assessment (a topic further elaborated by Buzick et al., 2023). A major advantage is that by using the assessments to enable test takers to solve problems, the assessment becomes a vehicle on route to solving the problem, and hence has the potential to promote learning (Gordon, 2020). DIRTy assessments can also be more culturally responsive. For example, students can have choice in which reading passages or other stimuli they respond to, and the choices can reflect the communities in which test takers live. Furthermore, the test delivery system can allow test takers to access translations of test material while taking the assessment, or even to assemble a different language version of the assessment.

Maybe it is time to learn from the fast-growing world of technology. We need to swallow hard and start with evidence-based interventions at the outset to provide models for teachers and then fix them in the wild (the much-reviled idea of flying a plane while repairing or improving them). This would lead, like in technologies, to numerous software updates (as we receive with our cars and phones) using the evidence from users to improve the impact of the software, track how prior users

have used and progressed when using the software, and de-implementing those features which hinder or have little evidence of impact.

Summary of mind frame #9

Technology has much to offer education, and with respect to assessments that serve learners, we can use technology to develop interactive assessments that are optimal for *each* learner, rather than a single assessment optimized for *most* learners. In addition to developing and delivering tests, technology can provide real-time feedback from assessments, and engage learners in understanding their assessment performance and what to do next. Technology has the potential to allow assessments to foster engagement in the testing process, and to be fully aligned with and integrated into instruction.

Mind Frame #10: Transform the purposes of assessment to move away from an overreliance on accountability to more continuous assessments used in learners' acquisition of understanding, motivation for learning, collaboration, and problem-posing settings.

A key theme of this chapter is assessments can be used more effectively to support student learning when we focus on the ways of thinking or the mind frames of the teachers and students. The advent of AI opens new opportunities, and there is not much confidence among teachers and many students about the benefits of the current assessment and accountability regimes.

Sireci (2021) identified four reasons why there has been a loss of confidence in educational assessments: (a) measurement professionals are often hypocritical (i.e., we impose standards, but don't follow them); (b) we present a censured history of educational and psychological testing to ourselves and the public, but the public knows better; (c) we focus on what was important 100 years ago, rather than what is important today; and (d) we are entrenched in a culture of distrust (see also Baker & O'Neil, 2020). Such "psychometric paralysis" requires a new way of thinking about assessments and asks for a de-emphasis of norm-referenced competitiveness in educational testing, except in those rare instances where examinees actually are competing for a benefit. To regain trust, we need to: (a) engage with teachers and other educators to collaboratively develop tests and interpret test scores; (e) reconceptualize our notions of standardization to make tests more flexible to students' needs and funds of knowledge; and (f) design test score reports for

students that emphasize their strengths, rather than their weaknesses. Essentially, we need to reorient our practices to value students more than a score scale.

Most classroom assessments rarely move past estimates of achievement, and few consider the students' strategies of learning, their emotions before, during, and after a lesson(s), and so often, the skills to work alone or in groups are considered. There needs to be a transformation of classroom assessment purpose from annual, time-controlled accountability assessments to more continuous assessments used in learners' acquisition of understanding, motivation for learning, collaboration, and deep application of knowledge in problem-solving, communication, and authentic settings.

Traditional assessment has focused on accountability, promotion, certification, and competition (e.g., admissions). Focusing the assessment purpose more directly on student learning aims to improve student learning greatly. Indeed, suppose we can frame the assessment problem as providing the most appropriate assessment for a particular learner at a particular point in time to provide specific information to support their learning. In that case, we will have solved a much more important problem than accountability. Such student-targeted assessments will have more validity to understand what students know and where they need to go next. We do not need to stop instruction to assess students for accountability purposes; we need assessments as part of instruction to guide it. If we solve the problems of developing, administering, and interpreting valid assessments supporting student learning, aggregating the results from those assessments for accountability will be relatively easy. It is time for assessments to be supportive, not disruptive. It is time to make learners full partners in the assessment process so the results benefit them and place their needs above those of the policy makers.

Gordon and Rajagopalan (2016) noted that too much testing overemphasizes the status of a narrow range of cognitive functions in learners and neglects the affective and situative domains of human performance and the processes by which these functions and domains are engaged. They neglect the diverse contexts and perspectives born of different cultural experiences and cultural identities and the influence of these contexts, perspectives, and identities on human performance. They have privileged accountability, relative positioning and competition to the neglect of criterion-based judgments of competence. They have overly focused on knowing, knowing how to, and mastery of knowledge that is held to be objectively

true. But they have confidence that assessment "can be a powerful and dynamic tool for effecting real transformation in how we view and deliver education in our society today and in the future."

Concluding Comments

In this chapter, we presented 10 mind frames for teachers, students, test developers, and all others involved in educational assessment that will move educational assessments from the measurement "of learning" to "supporting learning." Assessment best supports student learning when it is fully integrated with it. Baker (2018) has argued that it is well nigh that we "connect assessments systematically to instruction. Most approaches to establish instructional sensitivity of assessments to elements in learning programs are primitive at best, but needed for 'personalized' learning approaches" (p. 140).

Assessment should be organized as part of the instructional process rather than be disruptive to teaching and learning. Coordinating curriculum, instruction, and assessment is often described as alignment, which Webb (1997) described as "the degree to which expectations [i.e., standards] and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do" (p. 4). Including assessment in instruction requires considering when student evidence of learning is needed, what form that evidence should take, and how the evidence will inform future learning and instruction. Such considerations require aligning the content of assessments with the instruction objectives and timing the assessment at key points to inform instruction (Martone & Sireci, 2009). Without a doubt, the focus on improvement through design and teacher and student interpretations is essential, as are the moral and desired modifications to practice that support varied learners where they are. Policy changes that address tools and models for teachers, useful repositories of assessment, and evidence-based interventions should be integrated. We have been extolling the virtues of systematic codesign of teaching, learning, curriculum, and assessment since before Ralph Tyler's seminal efforts (1958). Perhaps creating usable tools that support infrastructure needed in curriculum design, teaching, and learning assessment may be an approach worth exploring once again.

Challenges remain. One is discovering how evaluative thinkers interpret and make decisions based on test scores and how we can best scale up these ways of thinking. Such scaling up remains one of the greatest unsolved issues

in educational research and practice. We see the opposite in developing contemporary technology, emphasizing scaling in a "first-to-market" mentality. While initial tests and evaluations might be undertaken, many technologies are released and then iteratively improved by collecting evidence of how users interact with the app, continually releasing updates.

In the domain of teaching, scaling of innovation or research-based options is hampered in at least three ways. First, many interventions need to be evaluated over a long period, not a day, a week, or in a single grade level, or on only one topic. Such trials are not practicable in a fast-moving market. Second, we continually argue that "our context is unique" and introduce adaptations that can take the innovation out of an implementation so that it becomes similar to what we had been doing anyway (hence, the innovation is too rarely actually implemented). Third, we so often scale without evidence of value, but based on teacher word-of-mouth or top-down mandates.

Scaling-up issues aside, we continue to think that the mind frame shift is likely the most important catalyst for accelerating student learning at a broad level for all students. As the mind frames in this chapter illustrate, by engaging and empowering students in their learning, by making them feel comfortable with their mistakes and allowing them the freedom to explore them, and by thinking about how assessments can be developed and used differently, we will have progressed beyond 19th and 20th-century ways of thinking about tests, to using them to support student learning.

References

Absolum, M., Flockton, L., Hattie, J. A. C., Hipkins, R., Reid, I (2009). *Directions for Assessment in New Zealand: Developing students' assessment capabilities.* Ministry of Education, Wellington, NZ. http://assessment.tki.org.nz/Assessment-in-the-classroom/Directions-for-assessment-in-New-Zealand-DANZ-report

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* Longman.

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory into practice*, *48*(1), 12–19.

Baird, J. A. (2014). Teachers' views on assessment practices. *Assessment in Education: Principles, Policy & Practice*, *21*(4), 361–364.

Baker, E. L. (2004). *Principles for Scaling Up: Choosing, Measuring Effects, and Promoting the Widespread Use of Educational Innovation.* CSE Report 634. Center for Research on Evaluation Standards and Student Testing CRESST.

Baker, E. (2018). Design for assessment change. *European Journal of Education, 53*(2), 138–140.

Baker, E. L., & O'Neil, H. F. (2020). The assessment landscape in the United States: From then to the future. *Monitoring student achievement in the 21st-century: European policy perspectives and assessment strategies*, 51–61.

Bandura, A. (1986). *Social foundations of thought and action.* Englewood Cliffs.

Bandura, A. (1993). *Perceived self-efficacy in cognitive development and functioning. Educational Psychologist, 28*(2), 117–148.

Becker, A. (2016). Student-generated scoring rubrics: Examining their formative value for improving ESL students' writing performance. *Assessing Writing*, *29*, 15–24.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome).* Academic Press.

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90.

Blannin, J., Hattie, J. A. C., Wood, C., & Stubbs, P. (in review). Informing Professional Learning Interventions with Evidence-Based Analysis of Student Feedback: Implications for Software Use and Learning Clarity

Bloom, B. S. (1971). *Handbook on formative and summative evaluation of student learning.* McGraw-Hill.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, Handbook I: The cognitive domain.* David McKay Co.

Booth, N. (2017, July 9). *What is formative assessment, why hasn't it worked in schools, and how can we make it better in the classroom?* Impact. https://my.chartered.college/impact_article/what-is-formative-assessment-why-hasnt-it-worked-in-schools-and-how-can-we-make-it-better-in-the-classroom/

Buckley, J., Archibald, T., Hargraves, M., & Trochim, W. M. (2015). Defining and teaching evaluative thinking: Insights from research on critical thinking. *American Journal of Evaluation*, *36*(3), 375–388.

Buzick, H. M., Casabianca, J., & Gholson, M. L. (2023). Personalizing large-scale assessment in practice. *Educational Measurement: Issues and Practice*, https://doi.org/10.1111/emip.12551

Clinton, J. M., & Hattie, J. A. C. (2021). *Cognitive complexity of evaluator competencies.* Evaluation and Program Planning, 89, 1–8.

Clinton, J., & Hattie, J. (2024). Revisiting and Expanding Scriven's Fallacies About Formative and Summative Evaluation. *Journal of MultiDisciplinary Evaluation*, *20*(47), 13–23.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics, 132*(4), 1593–1640.

Fisher, D., Frey, N., Ortega, S., & Hattie, J. A. C. (2023). *Teaching students to drive their learning: A playbook on engagement and self-regulation.* Corwin.

Francis, E. M. (2022). *Deconstructing depth of knowledge*. Solution Tree.

Frey, N., Hattie, J. A. C., &, Fisher, D. (2018). *Developing Assessment Capable Learners*. Thousand Oaks, Corwin.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice*, *39*(3), 72–78.

Gordon, E. W., & Rajagopalan, K. (2015). *The testing and learning revolution.* Jossey-Bass.

Hattie, J. A. C. (2014). The Last of the 20th-Century Test Standards. *Educational Measurement: Issues & Practice*, *33*(4), 34–35.

Hattie, J. A. C. (2018). *Implementing, scaling up, and valuing expertise to develop worthwhile outcomes in schools*. William Walker Oration, Presented at the Annual Conference of the Australian Council for Educational Leaders, Sydney. ACEL Monograph #58. http://www.acel.org.au/acel/ACELWEB/Publications/ACEL_Monograph.aspx

Hattie, J. A. C. (2023). *Visible learning: The sequel: A synthesis of over 2,100 meta-analyses relating to achievement*. Routledge.

Hattie, J. A. C., Brown, G., Ward, L., Irving, E., & Keegan, P. (2006). Formative evaluation of an educational technology innovation: Developer's insights into assessment tools for teaching and learning. *Journal of Multidisciplinary Evaluation, 5*(3), 1–54.

Hattie, J. A. C., Clarke, S., Fisher, D., & Frey, N. (2021). *Collective student efficacy.* Corwin.

Hattie, J. A. C., O'Leary, T., Hattie, K., & Donoghue, G. (2025). *Great Learners by Design.* Corwin.

Hattie, J. A. C., & Purdie, N. (1998). The Solo model: Addressing fundamental measurement issues. In Dart, B., & Boulton-Lewis, G. M. (Eds.), *Teaching and learning in higher education*. Camberwell, Vic, Australian Council of Educational Research.

Hess, K. (2006). *Exploring cognitive demand in instruction and assessment.* National Center for the Improvement of Educational Assessment, Dover NH.

Kapur, M. (2008). Productive failure. *Cognition and instruction*, *26*(3), 379–424.

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science*, *40*, 651–672.

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist, 51*(2), 289–299.

Kapur, M. (2024). *Productive Failure: Unlocking Deeper Learning Through the Science of Failing.* Jossey-Bass

Koretz, D. (2017). *The testing charade: Pretending to make schools better.* University of Chicago Press.

Law, N., Alexander-Hollins, Smith, D., & Hattie, J. A. C. (2024). *10 Mindframes for the culture and climate of schools: Equity, Identities, and Belonging*. Corwin.

Lee, J., LeBaron Wallace, T., & Alkin, M. (2007). Using problem-based learning to train evaluators. *American Journal of Evaluation*, *28*(4), 536–545.

Linden, W. J. (2005). *Linear models for optimal test design*. Springer Science+ Business Media, Incorporated.

Lord, A. Theory of test scores.-, 1952. *Psychometric Monograph*, (7).

Luecht, R. M., & Hambleton, R. K. (2021). Item response theory: A historical perspective and brief introduction to applications. In *The History of Educational Measurement* (pp. 232–262). Routledge.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction, *Review of Educational Research 4*, 1332–1361.

Miller, R. L., King, J. A., Mark, M. M., & Caracelli, V. (2016). The oral history of evaluation: The professional development of Robert Stake. *American Journal of Evaluation*, *37*(2), 287–294.

Moeller, J., Brackett, M. A., Ivcevic, Z., & White, A. E. (2020). High school students' feelings: Discoveries from a large national survey and an experience sampling study. *Learning and Instruction*, *66*, 101301.

Oberauer, K. (2010). Declarative and procedural working memory: Common principles, common capacity limits?. *Psychologica Belgica*, *50*(3–4), 277–308.

Piaget, J. (1952). *The origins of intelligence*. New York: International University Press.

Plass, J. L., & Kalyuga, S. (2019). Four ways of considering emotion in cognitive load theory. *Educational Psychology Review*, *31*, 339–359.

Roland, J. (1958). On "knowing how" and "knowing that". *The Philosophical Review, 67*(3), 379–388.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67.

Ryle, G. (1945, January). Knowing how and knowing that: The presidential address. In *Proceedings of the Aristotelian society* (Vol. 46, pp. 1–16). Aristotelian Society, Wiley.

Ryle, G. (1949). *The concept of mind*. Penguin Books.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144.

Sagone, E., & De Caroli, M. E. (2014). Relationships between psychological well-being and resilience in middle and late adolescents. *Procedia-social and behavioral sciences*, *141*, 881–887.

Schellings, G. (2011). Applying learning strategy questionnaires: Problems and possibilities. *Metacognition and Learning*, *6*, 91–109.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler (ed.), *Perspective of Curriculum Evaluation*, American Educational Research Association (AERA), Monograph of Curriculum Evaluation, No. 1., Rand McNally.

Scriven, M. (1993). The nature of evaluation. *New Directions for Program Evaluation*, *58*, 5.

Sireci, S. G. (2020). "De-"Constructing" Test Validation," *Chinese/English Journal of Educational Measurement and Evaluation*, 1. 教育测量与评估双语季刊: https://www.ce-jeme.org/journal/vol1/iss1/3 https://doi.org/10.59863/CKHH8837

Sireci, S. G. (2020). Standardization and understandardization in educational assessment. *Educational Measurement: Issues and Practice, 39*(3), 100–105. https://doi.org/10.1111/emip.12377

Sireci, S. G. (2021). Valuing educational measurement. *Educational Measurement: Issues and Practice, 40*(1), 7-16. https://doi.org/10.1111/emip.1241

Sireci, S. G., Suárez-Alvárez, J., Zenisky, A. L., & Oliveri, M. E. (2024). Evolving educational testing to meet students' needs: Design-in-real-time assessment. *Educational Measurement: Issues and Practice*, *43*(4), 112–118.

Snow, R. E., & Lohman, D. F. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology, 26,* 347–376.

Surma, T., Vanhees, C., Wils, M., Nijlunsing, J., Crato, N., Hattie, J., Muijs, D., Rata, E., Wiliam, D., & Kirschner, P. A. (2025). *Developing Curriculum for Deep Thinking: The Knowledge Revival.* Springer

Thissen, D., & Steinberg, L. (2020). An intellectual history of parametric item response theory models in the twentieth century. *Chinese/English Journal of Educational Measurement and Evaluation| 教育测量与评估双语期刊*, *1*(1), 5.

Thurstone, L. L. (1925). *The fundamentals of statistics* (Vol. 4). Macmillan.

Traub, R. E., & Wolfe, R. G. (1981). Chapter 8: Latent Trait Theories and the Assessment of Educational Achievement. *Review of Research in Education*, *9*(1), 377–435.

Tyler, R. W. (1958). Curriculum organization. *Teachers College Record*, *59*(11), 105–125.

Vo, A. T., Schreiber, J. S., & Martin, A. (2018). Toward a conceptual understanding of evaluative thinking. *New Directions for Evaluation*, *2018*(158), 29–47.

Webb, N. (1997). *Criteria for alignment of expectations and assessments on mathematics and science education*. Research Monograph Number 6. CCSSO.

Webb, N. L. (2002). Depth-of-knowledge levels for four content areas. *Language Arts*, *28*(March), 1–9.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*(1), 7–25.

Wyatt-Smith, C., & Adie, L. (2021). The development of students' evaluative expertise: Enabling conditions for integrating criteria into pedagogic practice. *Journal of Curriculum Studies*, *53*(4), 399–419.

Zenisky, A. L., O'Donnell, F., & Hambleton, R. K. (in press). Reporting scores and other results. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th edition). Oxford University Press.

Dynamic Pedagogy: A Perspective for Integrating Curriculum, Instruction, and Assessment in the Service of Learning at the Classroom Level

Eleanor Armour-Thomas

This chapter[1] responds to the Gordon Commission Report on the Future of Assessment (2013), which argued that assessment must go beyond documenting what students have learned or achieved. It must also gather evidence of the processes students use to learn and improve their learning over time. The chapter contends that realizing this vision requires embedding assessment within a framework of Dynamic Pedagogy—an integrated model in which assessment, curriculum, and instruction are inseparable and mutually reinforcing, all working together to promote learning with understanding.

The chapter begins by defining and conceptualizing Dynamic Pedagogy, positioning learning with understanding as the central focus of the interdependent relationships among assessment, curriculum, and instruction. It then outlines how Dynamic Pedagogy can be operationalized at the classroom level, beginning with the identification of desired learning outcomes and clearly defined success criteria. This is followed by a discussion of three essential components for implementation: (1) teaching-learning transactions, (2) a learning-centered environment, and (3) a

1 This chapter builds upon concepts explored in earlier work (See Armour-Thomas, E., & Gordon, E. W. (2025). *Principles of Dynamic Pedagogy: An integrative model of curriculum, instruction, and assessment for prospective* and in-service teachers, Routledge), expanding the analysis to include new contexts and applications.

three-phase instructional structure—Pre-active, Interactive, and Post-active—that collectively support both service and deep learning.

The chapter further examines the multiple functions of a classroom assessment system and the purposes it serves: generating evidence of learning as it occurs and determining the status of what has been learned. It concludes by situating Dynamic Pedagogy within a broader assessment system and proposing a research and development initiative to evaluate the model's effectiveness in enabling learning with understanding at the classroom level.

Introduction

A common theme in the educational assessment literature is its critical relationship to learning and its improvement. There are at least two key issues that must be addressed in the practice of assessment within educational settings. First, as recommended by the Gordon Commission on the Future of Assessment, assessment should not be limited to generating evidence about students' current levels of achievement. It should also provide insights into the processes students use to learn and to improve their learning over time. Second, assessment decisions alone cannot meaningfully inform or support learning. This is because assessment is inextricably linked to other core components of pedagogy—namely, curriculum and instruction (Armour-Thomas & Gordon, 2013; Farenga et al., 2002; Gordon, 1999; Furtak et al., 2016; Popham, 2008; Shepard, 2021; Tomlinson & Moon, 2013; Wiliam & Leahy, 2015). How, then, might assessment in the service of learning be conceptualized and implemented, given its inseparable relationship with curriculum and instruction? This chapter explores these issues through the lens of Dynamic Pedagogy—an approach to teaching and learning in which learning lies at the center of three interdependent domains: assessment, curriculum, and instruction. These elements function in dynamic interaction to promote meaningful learning and its continuous improvement.

Definition and Conceptualization of Dynamic Pedagogy

Dynamic Pedagogy is an approach to teaching and learning in the classroom where the primary purpose of pedagogical processes of curriculum, instruction, and assessment is to enable learners to learn with understanding. Learning with understanding can be defined as an individual's capacity to grasp the conceptual meaning of key ideas within a discipline, recognize their significance, and apply them flexibly in varied contexts. It extends beyond memorization or procedural

proficiency to include the development of deep, transferable knowledge that can be used to solve novel problems, make informed decisions, and generate new insights about what it means to learn with understanding.

This definition is grounded in social-constructivist perspectives on how people learn (e.g., Bransford et al.,2002) which emphasize that learners actively construct meaning based on their prior knowledge and experiences. As Schmidt and Marzano (2015) and Bailey and Pransky (2014) argue, the integration of new knowledge is shaped by what learners already know, making the activation and refinement of prior understanding essential to deeper learning.

Learning with understanding also involves the consolidation of new knowledge into coherent mental models, as Greene (2008) suggests, and its transferability to unfamiliar situations—a cornerstone of expert thinking, as described by Bransford and Stein (1993) and Perkins and Salomon (2012). In this way, learners are not only able to recall facts but also to reconstruct knowledge in new contexts, which is a hallmark of meaningful learning.

Moreover, this form of learning engages students in higher-order thinking, such as analysis, evaluation, and creation, as emphasized by Fosnot (2005). It encourages self-regulated learning (Zimmerman & Schunk, 2013), where students set goals, monitor their own progress, and reflect on outcomes. These metacognitive processes foster greater autonomy and ownership over their learning.

Finally, learning with understanding is closely related to what Gordon (2007) terms intellective competence—the ability to use one's knowledge base flexibly, insightfully, and strategically in the face of cognitive challenges.

For learners to learn with understanding, they must be able to demonstrate that they can activate their prior knowledge and use it to construct and consolidate new knowledge, as well as transfer it to other contexts. Classroom teachers support students in learning to learn with understanding using an approach to teaching and learning called "Dynamic Pedagogy." In this definition, "Pedagogy" denotes the interdependency of curriculum, instruction, and assessment processes with learning as its collective focus. The term "dynamic" emphasizes the fluidity of the interdependent relationships among curriculum, instruction, and assessment with learning processes for the purpose of enabling students to learn with understanding.

A logical coherence exists among the three pedagogical processes, each having a complementary and functional relationship with learning processes. For instance, the assessment and curriculum processes are coherently related in that the choice of which assessment probe to use in the inquiry about a learner's engagement in a discipline-specific problem-solving task depends in part on the level and complexity of the curriculum task and its attendant cognitive, metacognitive, and motivational demands on the learner. Hence, both will influence learning. Furthermore, interpretations of results from analysis of assessment data can guide teachers in adjusting instruction and support learners in making necessary changes to their learning strategies, thus demonstrating the inseparable nature of assessment, instruction, and learning. Similarly, instructional and curriculum adaptations informed by interpretations of results about learners' performance from curriculum-embedded assessment tasks are yet another illustration of the interdependent relationships of pedagogical and learning processes.

In this chapter four concepts of learning processes are selected for learners' responsiveness to individual components of Dynamic Pedagogy (curriculum, assessment and instruction as well as the interdependence among them): cognitive processes (Anderson & Krathwohl (2001); executive processes (Flavell, 1979; Schunk & Zimmerman, 2016); motivational orientation (Schunk, 2016; Solomon & Anderman, 2017); sensory response modalities (Mayer, 2008; Nesbit & Adesope, 2006; and Sternberg, 1998). This chapter argues that learners utilize these concepts when prompted to draw on their prior knowledge to construct and consolidate new knowledge and to transfer newly acquired knowledge to other contexts. For a more comprehensive discussion of these ideas, see Armour-Thomas (2017), Armour-Thomas and Gordon (2019), and Armour-Thomas and Gordon (2025).

While there are substantive research findings that attest to the relationship between each of these pedagogical processes of teaching and learning, this chapter argues that it is the interplay among them, with learning as the central focus, that most significantly impacts learning more than any single pedagogical process. The notion of pedagogical interdependence and its relationship with learning processes traces back to Tyler's foundational work (1949) and has been reinforced by subsequent scholarship (e.g., Farenga et al., 2002; Gordon, 1999; Pellegrino et al., 2014; Shepard, 2021; Tomlinson & Moon, 2013 & Wiliam, 2011).

Operationalization of Dynamic Pedagogy in the Classroom

Learning Targets and its Criteria for Success

The operationalization of Dynamic Pedagogy in the classroom begins with the identification of learning outcomes students are expected to achieve by the end of lessons of a curriculum unit along with clearly defined success criteria. Where learners are in relation to the outcomes expected of them, as well as the progress they would need to demonstrate on their way to mastery, are also considerations in the operationalization of Dynamic Pedagogy. The learning targets and their success criteria guide both the teacher's and the learner's actions to support learning with understanding throughout lessons of a curriculum unit.

Teachers reflect on questions such as:

1. What curriculum-embedded assessments will I use to ascertain where students are in relation to the learning target?
2. What curriculum-embedded instruction and assessment will I use to help students make progress toward the learning target?
3. What curriculum-embedded assessments will I use to ascertain whether students met the learning target?

Learners consider questions like:

1. Do I understand the learning target and the criteria for success?
2. Where am I in relation to the learning target?
3. What do I need to do to make progress toward the learning target and meet it?

Research over the last two decades has affirmed the importance of these steps in the learning process (Brookhart, 2024; Moss & Brookhart, 2012; Wiliam, 2011; Wiliam & Leahy, 2015). Three additional factors are essential for operationalizing Dynamic Pedagogy in the classroom: teaching-learning transactions, a learning-centered environment, and a three-phase structure for linking the functional relationships of the interdependent pedagogical processes with learning processes.

Teaching-Learning Transactions

Teaching-learning transactions refer to the reciprocal relationship between teachers and students within instructional arrangements in the classroom (e.g., whole group, small group, dyads, and one-to-one relationships between peers or teacher and learner). It is within these arrangements that mediational teaching-learning transactions unfold and the mechanism for learning with understanding lies. It is here that dialogic exchanges about curriculum-embedded instruction and instruction-embedded assessments occur, and where learners, in collaboration with their peers, are encouraged by their teacher to make their cognitive and metacognitive thinking audible and visible as they build on their prior knowledge to construct, consolidate, and transfer new knowledge to other contexts. It is also here that the teacher helps students to broaden their cognitive and motivational schemas through curriculum-embedded instruction and assessment while being responsive to their preferred modalities of expressing what they know and can do. Both teacher and student thus share responsibility for fostering learning with understanding.

Learning-centered environment

The learning environment of Dynamic Pedagogy encompasses not only the physical space but also the emotional and social atmosphere of the classroom. It is thus a learning-centered environment where all learners are socially connected with their peers and teachers to engage in new or challenging activities with them. Characteristics such as personalization, trustworthiness, rapport, empathy, and care are essential for positive teaching-learning transactions, as these elements foster student engagement, confidence, and a sense of agency in and ownership of their learning.

Effective learning-centered environments that foster learning with understanding are also shaped by administrative routines (combinations that maximize active participation), the appropriate integration of pedagogical processes (which support goal-directed learning), and pacing (which regulates opportunities for student expression and interaction with the teacher and their peers). An elaboration of what Dynamic Pedagogy looks like in practice is provided below.

A Learning Focus of Pedagogical Functions

A three-phase structure is used to operationalize Dynamic Pedagogy at the classroom level—preactive, interactive, and postactive—as informed by the work of Jackson (1968) and Artzt & Armour-Thomas (2002). This structure describes a sequence of metacognitive thinking of planning, monitoring, evaluating, and revising lessons of a curriculum unit. The focus of these higher-order processes for the teacher is on the relationships of the interdependent components of the pedagogical components of curriculum, instruction and assessment with learning processes to: support learning with understanding while it is occurring and to determine how much learning is achieved at the end of lessons of a curriculum unit.

1. The Preactive Phase

In the preactive phase, the teacher engages in thoughtful planning of lessons within a curriculum unit (e.g., activities about the integration of pedagogical components of curriculum, instruction and its relationship with students' needs, assets and interests; the teaching-learning transactions where these activities will unfold; and the features of the classroom environment conducive to learning with understanding).

Key teacher metacognitive thoughts in the preactive phase include:

- The learning targets and their success criteria in a domain-specific area that students are expected to demonstrate by the end of lessons in a curriculum unit.
- The interdependent curriculum and assessment strategies that would be needed to diagnose where students are in relation to where they need to be at the end of a lesson as well as the combinations of curriculum, instruction, and assessment tasks that would be needed to support students' progress toward the learning targets of lessons within a curriculum unit.
- The misconceptions that may be uncovered from the diagnostic prompts and the combinations of curriculum, instruction, and assessment processes that would be required to address them.
- The differentiated entry points into the lesson that would be needed to help students build on their prior knowledge to construct new knowledge, to consolidate and transfer it to new contexts.

The modalities for representing curriculum, instruction, and assessment tasks, as well as the allowable modalities (e.g., oral, written, visual, or kinesthetic) for students to demonstrate what they know and can do in response to curriculum-embedded assessments.

2. The Interactive Phase

During the interactive phase, the teacher enacts the lesson plan, which incorporates various combinations of curriculum, instruction, and assessment, with learning processes as the focus and learning with understanding as its primary purpose. Diagnostic and appraisal assessments are also used to gather evidence about student performance, about how well students are making progress toward the learning targets and their criteria.

Key teacher metacognitive thoughts in the interactive phase include:

- The diagnosis of students' readiness for new learning by administering curriculum-embedded assessments to ascertain their prior knowledge and skills and their underlying cognitive processes. How well does the evidence from diagnostic assessments provide information about how close students' performances are to the learning targets and their success criteria?
- The appraisal of students' progress toward new learning by administering instruction and curriculum-embedded assessments to ascertain how they use their prior knowledge to construct new knowledge and skills and their underlying cognitive processes. Such assessments are also used to ascertain how students consolidate and transfer new knowledge and skills, as well as their underlying cognitive processes. Similar to the metacognitive thinking above, the teacher thinks about how well the evidence from appraisal assessments provides information about how close to the learning targets and success criteria are students' performances.
- The scaffolded learning experiences that enable students to engage with complex tasks while receiving timely feedback to refine their thinking and performance.
- Learners' choice of the use of multiple assessment modalities so that students can demonstrate what they know and can do in ways that align with their intellective strengths, interests, and needs.

- The use of inferences from analysis of assessment results to inform adjustments in subsequent relationships between interdependent pedagogical components of curriculum, instruction and assessment and learning processes to inform learning with understanding.
- The use of inferences from analysis of results of curriculum-embedded assessments to provide feedback to students about how well they are making progress toward the learning targets of the lesson.

3. The Postactive Phase

In the postactive phase, the teacher self-evaluates how well curriculum-embedded assessments generate evidence about the status of students' achievement of learning targets of the lesson. The teacher also self-evaluates about the degree of students' mastery of the learning targets at the end of the curriculum unit.

Key teacher self-evaluative thoughts in the postactive phase to ascertain the status of learning achieved at the end of a lesson and at the end of a curriculum unit include:

- The design of curriculum-embedded assessments with various formats and modalities of representation of items that are likely to generate evidence of the degree of learning achieved at the end of a lesson and the curriculum unit.
- Students' choice of modalities in responding to curriculum-embedded assessments to demonstrate what they know and can do at the end of a lesson and the curriculum unit
- The inferences made from results of curriculum-embedded assessments to inform revisions in subsequent relationships of pedagogical components of curriculum, instruction and assessment and learning processes to inform learning with understanding. Such revisions can be made in subsequent lessons of a new curriculum unit.

Assessment System at the Classroom Level

At the classroom level, an effective assessment system for Dynamic Pedagogy generates meaningful information about learning before, during, and after pedagogical interventions. These interventions reflect the synchrony of assessment, curriculum, and instruction—unified in their collective focus on supporting learning with understanding. Specifically, an assessment system fulfills three primary functions:

1. **Diagnosis**–identifying where learners currently are in relation to the outcomes of learning set for the end of a curriculum unit.
2. **Appraisal**–monitoring learners' progress toward mastery of those outcomes of learning.
3. **Evaluation**–determining the level of mastery learners have achieved by the end of a lesson as well as at the end of a curriculum unit.

These functions serve two overarching purposes:

4. To **generate evidence that learning is occurring** as students engage in lesson activities within a curriculum unit; and
5. To **generate evidence about the status of learning achieved** after the curriculum unit concludes.

Regardless of the assessment's purpose, the process typically follows a consistent sequence:

(a) Generate data on what students know and can do;

(b) Analyze assessment results;

(c) Draw inferences from the analysis of assessment results;

(d) Use inferences from the analysis of assessment results to inform next steps for the teacher and the student.

Purpose 1: Evidence to Inform Learning while it is Occurring.
When the focus of assessment is to seek evidence that learning is occurring during dynamic pedagogical interactions between the teacher and students, the following five design principles of assessment are proposed for formative purposes:

1. **Open-ended prompts** that allow students to choose their preferred modality of expression (e.g., visual, oral, written) to demonstrate what they know and can do as they make progress toward the learning targets and their success criteria.
2. **Complexity tracking** of cognitive and metacognitive processes as students engage with knowledge-based tasks of lessons within a curriculum unit.
3. **Observation of learning behaviors**, capturing nature and quality of student engagement in dynamic pedagogical activities.
4. **Rubrics** that yield qualitative and quantitative evidence of how students are making progress toward the learning targets and the success criteria of a curriculum unit.
5. **Inferences from assessment results** that teachers use as feedback to make adjustments in one or more combinations of interdependent pedagogical processes of curriculum, instruction, and assessment to support student learning with understanding.

Purpose 2: Evidence to inform the Status of Learning Achieved
When the purpose shifts to generating evidence about the **quality and quantity of learning achieved** after a curriculum unit ends, the design principles of assessment remain closely aligned with those for formative purposes, with slight variations in emphasis:

1. **Open-ended items** that give learners a choice of their preferred modalities of expression to demonstrate what they have learned at the end of a curriculum unit.
2. **Cognitively rich prompts** are indicative of a full range of cognitive processes, underlying knowledge-based tasks to ascertain what and how much students have learned at the end of a curriculum unit.
3. **Rubrics** that yield both qualitative and quantitative evidence of how much students have learned at the end of a curriculum unit.
4. **Inferences from assessment results** inform subsequent decision-making and actions by the teacher to improve student learning with understanding through dynamic pedagogical activities.

Situating Dynamic Pedagogy Within a Larger Educational System

While the conceptual and methodological foundations of Dynamic Pedagogy are anchored at the classroom level, it is essential that the model be situated within and supported by broader systemic structures. As research continues to illuminate how students learn and the conditions that foster learning with understanding over time, it becomes increasingly clear that school, district, and state-level policies and practices must play a role in ensuring that Dynamic Pedagogy does not operate in a vacuum.

For example, if teachers are to engage with and adopt the principles of Dynamic Pedagogy meaningfully, they must be provided with sustained professional development and resource support. Such enhancements should be coordinated by school and district personnel who are committed to advancing learner-centered practices throughout the entire educational system.

Although conventional large-scale assessments at district and state levels share certain design features with the assessment system of Dynamic Pedagogy—such as the incorporation of multiple cognitive processes and allowances for diverse response modalities, more systemic coherence is needed. Policymakers, assessment developers and curriculum and instruction experts must collaborate to align the design and implementation of classroom-based assessments with interim and end-of-year large-scale assessments. This alignment is crucial for building a balanced and integrated assessment system, one that is grounded in both theory and evidence about how learners develop understanding within meaningful contexts (e.g., Darling-Hammond et al., 2013; Evans & Marion, 2024; and Marion, Pellegrino, & Berman, 2024).

When the core principles and features of a classroom-based Dynamic Pedagogy assessment system are aligned with those of assessments beyond the classroom, both students and teachers stand to benefit. For students, sustained engagement with curriculum-embedded assessments designed to promote learning with understanding enhances their preparedness, not only for classroom-based summative assessments but also for standardized assessments administered at district and state levels. While this hypothesis warrants empirical investigation, it represents a promising direction for future research and practice.

For teachers, the thoughtful use of large-scale assessment data—when such data are meaningfully connected to curriculum and instruction—can enrich pedagogical decision-making. This alignment strengthens teachers' capacity to design meaningful curriculum and instruction embedded assessments that support students' deep learning and long-term academic growth.

Future Directions for Dynamic Pedagogy

It is one thing to make a conceptual argument for the perspective of Dynamic Pedagogy to inform learning with understanding at the classroom level, but quite another to demonstrate its efficacy in practice. Further research is needed to evaluate the effectiveness of its claims in practice. Suggested areas of inquiry include:

1. What validity criteria inform the design and use of assessments about the Dynamic Pedagogy model?
2. What are the essential components of a professional development program that support the planning, implementation, and evaluation of Dynamic Pedagogy?
3. What kinds of infrastructures are required to plan, implement, and evaluate Dynamic Pedagogy effectively?
4. How can stakeholders participate meaningfully in the development and sustainability of Dynamic Pedagogy interventions?
5. How can technology be used to enhance the planning, implementation, and evaluation of Dynamic Pedagogy?

Conclusion

This chapter proposed an approach to teaching and learning defined as Dynamic Pedagogy, where curriculum, assessment, and instruction are interdependent pedagogical processes with the shared goal of fostering student learning with understanding. The operationalization of Dynamic Pedagogy calls for a shared responsibility between teachers and students to engage in teaching-learning transactions situated within a learning-centered environment that fosters students' confidence, agency, and ownership in their own learning. The chapter proposed design principles of an assessment system oriented toward generating evidence of Dynamic Pedagogy's influence on student learning with understanding. Finally, it recommended a research and development agenda to validate and refine the practices associated with Dynamic Pedagogy, ensuring its positive impact on student learning and understanding at the classroom level.

References

Armour-Thomas, E., & Gordon, E. W. (2025). *Principles of Dynamic Pedagogy: An integrative model of curriculum, instruction, and assessment for prospective and in-service teachers,* Routledge.

Armour-Thomas, E. (2017). Dynamic Pedagogy: An Integrative Model of Assessment, Curriculum, and Instruction in the Service of Learning. *Journal Leadership and Policy Studies* 1, 22–37

Armour-Thomas, E. (2019). Formative assessment: An approach to growing student learning while it's occurring. In E. Armour-Thomas, M. Wade Boykin, & E. W. Gordon (Eds.), *Human variance and assessment for learning* (pp. 345–366). Third World Press Foundation.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* Allyn & Bacon.

Artzt, A. F., & Armour-Thomas, E. (2002). *Becoming a reflective mathematics teacher: A guide for observations and self-assessment.* Lawrence Erlbaum Associates.

Bailey, F., & Pransky, K. (2014). *Memory at work in the classroom: Strategies to help underachieving students.* ASCD.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How People Learn: Brain, Mind, Experience, and School* (Expanded ed.). Washington, DC: National Academy Press.

Bransford, J. D., & Stein B. S. (1993). *The IDEAL Problem Solver* (2nd ed.). Freeman.

Brookhart, S. M. (2024). *Classroom Assessment Essentials*. ASCD.

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., & Steele, C. M. (2013). *Criteria for high-quality assessment.* Stanford Center for Opportunity Policy in Education.

Evans, C. M., & Marion, S. F. (2024). Building assessment systems that support instructional coherence. *Educational Measurement: Issues and Practice, 43*(1), 12–24.

Farenga, S., Joyce, B. A., & Ness, D. (2002). In R. Bybee (Ed.), *Learning science and the science of learning* (pp. 51–62). National Science Teacher Association Press.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34, 906–911.

Fosnot, C. (Ed.). (2005). *Constructivism: Theory, perspectives, and practice* (2nd ed.). New York, NY: Teachers College Press.

Furtak, E. M., Kiemer, K., Circi, R. K., Swanson, R., de Leon, V., Morrison, D., & Heredia, S. (2016). Teachers' formative assessment abilities and their relationship to student learning. Findings from a four-year intervention study. *Instructional Science*, 44(3), 267–291.

The Gordon Commission on the Future of Assessment in Education. (2013). To assess, to teach, to learn: *A vision for the future of assessment* (Technical Report). Educational Testing Service.

Gordon, E. W. (1999). *Education Justice: A View from the Back of the Bus.* Teachers College Press.

Gordon, E. W. (2007). Intellective Competence: The universal currency in technologically advanced societies. In E. W. Gordon and B. Bridglall, *Affirmative Development: Cultivating Academic Ability* (pp. 3–16). New York: Rowman & Littlefield.

Greene, R. L. (2008). Repetition and spacing effects. In J. Byrne (Ed.), *Learning and memory* (pp. 65–78). Offord, England: Elsevier.

Jackson, P. W. (1968). *Life in classrooms.* Holt, Rinehart, and Winston.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge University Press.

Marion, S. F., Pellegrino, J. W., & Berman, A. I. (Eds.). (2024). *Reimagining balanced assessment systems.* National Academy of Education.

Mayer, R. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psychologist*, 63(8), 757–769.

Moss, C. M., & Brookhart, C. M. (2012). *Learning targets: Helping students with their aim for understanding in today's lesson.* ASCD.

Nesbit, J., & Adesope, O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research, 76*(3), 413–448.

Pellegrino, J. W., Wilson, M., Koenig, J., & Beatty, A. (Eds.). (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: National Academy Press.

Popham, W. J. (2008). *Transformative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Perkins, D., & Salomon, G. (2012). Knowledge to go: A motivational and dispositional view of transfer. *Educational Psychologist, 47*(3), 248–258.

Schunk, D. (2016). *Learning theories: An educational perspective* (7th ed.). Boston, MA: Pearson.

Schunk, D., & Zimmerman, B. (2013). Self-regulation and learning. In W. Reynolds, G. Miller, & I. Weiner (Eds.). *Handbook of psychology* (Vol. 7, 2nd ed., pp. 49–69). Wiley.

Shepard, L. A. (2021). Ambitious teaching and equitable assessment: A vision for prioritizing learning, not testing. *American Educator, 45*(3), 28–48.

Solomon, H., & Anderman, E. (2017). Learning with motivation. In R. Mayer & P. Alexander (Eds.). *Handbook of research on learning and instruction* (2nd ed., pp. 258–282). Routledge.

Schmidt, R., & Marzano, R. (2015). *Recording and representing knowledge: Classroom techniques to help students accurately organize and summarize content*. West Palm Beach, FL: Learning Sciences International.

Sternberg, R. (1998). Principles for teaching successful intelligence. *Educational Psychologist, 33*(2/3), 65–72.

Tomlinson, C. A., & Moon, T. (2013). *Assessment and student success in a differentiated classroom*. ASCD.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. University of Chicago Press.

Wiliam, D. (2011). *Embedded formative assessment*. Solution Tree Press.

Wiliam, D., & Leahy, S. (2015). *Embedding formative assessment: Practical techniques for K–12 classrooms*. Learning Science International.

Assessment as a Pillar of Pedagogy in Support of Learning in AP Research and Mathematics Education Courses

Eleanor Armour-Thomas, Jacqueline Darvin, and Gerunda B. Hughes

Abstract

Assessment serves as a fundamental pillar of pedagogy, functioning in synchrony with curriculum and instruction to support student learning at the classroom level. This perspective aligns with Armour-Thomas and Gordon's (2013) concept of Dynamic Pedagogy, where these three pedagogical components are interdependent and reciprocally related with support for learning as its collective purpose.

Examples from AP Research and Mathematics Education courses illustrate how assessment is inseparable from curriculum and instruction, highlighting key takeaways: ascertaining where learners are in their current learning in relation to where they need to be in terms of learning goals, appraising their progress toward the learning goals, and figuring out next steps in their learning toward mastery of the learning goals. When these pillars interact effectively, learning becomes visible through students' demonstration of knowledge and skills in varied ways. Additionally, teachers and students can use the results of analysis of assessment data as actionable feedback that inform next steps in the interdependent pedagogical processes of instruction, curriculum, assessment with learning and its improvement as its focus.

This chapter references the Handbook principles of assessment in the service of learning such as transparency, assessment design, and feedback. Moving forward, empirical research is needed to explore assessment's role as a pillar of pedagogy and its enablement of learning and its improvement.

Changing conceptions of how students learn, along with educational standards that call for all students to develop deeper learning 21st-century competencies, have heightened interest in assessments that are responsive to these changes. Over the years, policy makers, educators and researchers have given much attention to the power of classroom assessments and their relationship to student learning. Black and Wiliam (1998) provided compelling results in their review of empirical studies about the positive relationship between classroom formative assessment and student learning. Other research studies have reported similar findings about the positive relationship between classroom assessment and student learning (Stiggins & Chappuis (2012); Kingston and Nash, 2011; Hughes, 2010; Furtak et al., (2016); Popham, (2008); Shepard, (2021); Johnson et al., (2019).

We use the metaphor of a "pillar" to emphasize the significance of classroom assessment in supporting learning in the same way that a pillar supports a building. We also make the claim that what gives it its foundational support for learning is its interdependence with two components of teaching–curriculum and instruction. This claim is not new since others have recognized the interdependence of the different components of the teaching, learning and assessment processes (Armour-Thomas & Gordon, 2013; Black, 2018; Farenga, Joyce, & Ness, 2002; Gordon, 2020; Chatterji, 2012; Heritage, 2007; Wiliam & Thompson, 2007; Tomlinson & Moon, 2013; Tyler, 1949).

The next section of this chapter elaborates on the conception of assessment as a pillar of pedagogy followed by two examples from courses in AP Research and Mathematics Education that demonstrate how assessment as a pillar of pedagogy supports teaching and learning in the classroom. Also, both examples refer to the *Principles of Assessment in the Service of Learning.*

Assessment as a Pillar of Pedagogy

The conception of assessment as a pillar of pedagogy at the classroom level functions as a data-gathering procedure that is designed and used to generate evidence of its interdependent relationships and functions with two other pedagogical processes, curriculum and instruction, with supporting learning as their primary purpose and focus.

Firstly, the interdependence between assessment and curriculum and its relationship to learning is an example of assessment as a pillar of pedagogy to support learning. Before a lesson or curricular unit begins, the teacher uses a variety of diagnostic tools to ascertain students' readiness for the expected outcomes of the lesson or curricular unit. Interviews or questionnaires may be used by the teacher to gather information about students' preferred ways of demonstrating what they have learned that would be relevant for the expected mastery of the goals and objectives of the lesson or curricular unit. Their cultural, social ways of knowing, their perceptions and dispositions about learning, and the environments most conducive to them learning well are some of the information generated from curriculum-embedded assessment tools that can be useful for teachers prior to the start of the lesson or curricular unit.

Another example of assessment as a pillar of pedagogy is its interdependence with instruction and curriculum that occurs during instruction. Instruction-embedded assessment can be used to appraise how well students are learning with an understanding of what is being taught, and to what extent students are making progress toward mastery of the goals and objectives of the lesson or curricular unit. Such questions include oral and written questions, classroom and homework assignments and quizzes. Interpretation of results from these appraisal assessments during this phase of the teaching-learning process can yield feedback for both the teacher and the learner. For the teacher, feedback can be used to adjust curriculum and/or instruction in subsequent instruction. For learners, feedback can be used to inform them of the next steps. Such actions can include seeking specific help from resources, peers, and/or the teacher to address problematic aspects of their recent performance that would result in improvement in their performance. If no improvement is needed, learners can use feedback to strength their learning in readiness for moving toward meeting the goals and objectives of the lesson or curriculum unit.

A third example of assessment as a pillar of pedagogy is its interdependence with curriculum at the end of the lesson or curricular unit. Here, curriculum-embedded and instruction-embedded assessment tools are evaluative of mastery of the goals and objectives of the lesson or curricular unit. Think-Pair Share, exit tickets, quick writes, and reflections are examples of these end-of-lesson assessments and demonstrations or presentations, projects or essays, and unit tests are examples of end-of-unit assessments.

This conception of assessment as a pillar of pedagogy bears some similarity to Armour-Thomas and Gordon's notion of *Dynamic Pedagogy* (Armour-Thomas & Gordon, 2013) in which assessment, curriculum and instruction are reciprocally related with learning as its collective focus and purpose. Figure 1 illustrates the dynamic interdependence of these pedagogical processes and their alignment with learning processes.

Figure 1
Dynamic Pedagogy Model

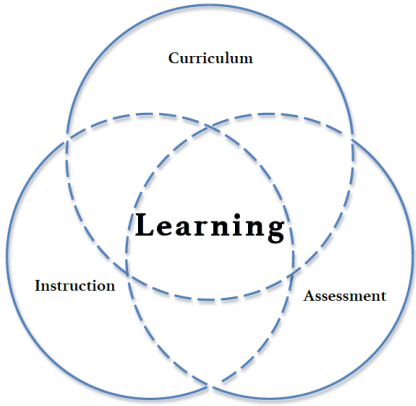


Figure 1. Interlocking circles indicate the interdependence of assessment, curriculum and instruction and the jagged lines are intended to depict the dynamic interaction among these three areas with learning as the focus (reproduced from a paper written for the Gordon Commission on the Future of Assessment in Education: Armour-Thomas & Gordon, 2013 Assessment as a dynamic component of Pedagogy).

What follows next are examples of assessment as a pillar of pedagogy in support of learning in AP Research and Mathematics Education courses.

Example One: AP Research Scaffolded Activities

Assessment as a pillar of pedagogy supports teaching and learning in the classroom and joins with the two other pillars of pedagogy, curriculum and instruction. When teachers reimagine and design assessments and learning tasks leading up to assessments that adhere to the *Principles for Assessment in the Service of Learning,* they make their feedback to learners, teachers, administrators, and parents more transparent, timely, and actionable. This can be accomplished by employing ongoing classroom formative assessments that are exploratory and reflective in nature and include a series of scaffolded learning opportunities and performance tasks for students to engage in as part of their ongoing learning processes. One such example where assessment undoubtedly can be in the service of learning for students, teachers, families, and school administrators is the AP Research course and assessment tasks that are administered by the College Board and ETS.

According to the College Board, 2024

> *AP Research, the second course in the AP Capstone experience, allows students to deeply explore an academic topic, problem, issue, or idea of individual interest. Students design, plan, and implement a yearlong investigation to address a research question. Through this inquiry, they further the skills they acquired in the AP Seminar course by learning research methodology, employing ethical research practices, and accessing, analyzing, and synthesizing information. Students reflect on their skill development, document their processes, and curate the artifacts of their scholarly work through a process and reflection portfolio. The course culminates in an academic paper of 4,000–5,000 words (accompanied by a performance, exhibit, or product where applicable) and a presentation with an oral defense* (p. 10).

By design, AP Research is an interdisciplinary course in which students, who are mostly 11th graders, conduct year-long research studies on their chosen topics. The course culminates with an academic paper and a presentation and oral defense to a panel of teachers and educational leaders. AP Research and

other inquiry-based learning scenarios rely heavily on the peer-review process for students to give and receive actionable feedback. Students are encouraged to seek feedback from their peers and expert advisors or mentors with backgrounds in the disciplines of study.

A feedback-focused collaboration began in 2020 with two AP Research teachers from East Hampton High School in Long Island, New York, meeting with a professor from Queens College, CUNY, to discuss a challenge they were having with an upcoming AP Research lesson. The lesson instructed high school students to provide feedback on each other's AP Research paper drafts, using a highlighting activity and annotated rubric. While some students were able to provide effective peer feedback, the AP Research teacher, Michael, indicated to the professor that others had difficulty commenting on the bottom two rows of the rubric that deal specifically with the communication of the student's ideas through the organization, design elements, grammar, style, word choice, and the proper citing and attributing of sources. The professor suggested that Michael should encourage students to focus on only one rubric component at a time. Starting small, with a primary focus, is less daunting to students and fosters better peer feedback. This strategy has since been applied in the AP Research classroom, resulting in more comprehensive student-to-student peer feedback, particularly on the two rows of the rubric with which students have the most difficulty. After applying the strategy, Michael commented, "What was once an intimidating task for many students is now manageable due to the activity being broken down into digestible pieces. My students are now able to provide constructive criticism to their peers and are no longer limited due to rubric fatigue."

Other examples of scaffolded classroom assessment activities that support student learning occur when the AP Research teacher takes a creative and recursive peer feedback approach with his high school students and uses different interactions of "speed dating" peer feedback opportunities throughout the course. When students are preparing their inquiry research proposals in the fall, they create digital posters and present an "elevator pitch" to their peers for feedback and to keep up with their presentation skills throughout the year. This is repeated at distinct stages, as students add more elements to their posters, such as their Research Questions, Methodology, etc. Michael instructs the AP Research students to support their peers, as they improve their research proposals, by providing

positive feedback via pink Post-it notes to help their peers see where they are doing things right and provide constructive feedback via yellow Post-it notes to show where their peers can improve.

He gives the students explicit instruction on how to provide feedback to their peers, including: When you post, be respectful. When you receive feedback, keep an open mind. Your reviewer is not attacking you. You are both attacking your research project for improvement. Make changes when they are fresh (start addressing feedback now!!). Additionally, Michael introduces key terms for the student researchers to use with one another while providing each other with feedback, including: 1. Broad: Generic, vague, covering too many subjects or areas, 2. Clear: Easy to understand, not confusing 3. Focus on the specific concept that is being addressed, 4. Narrow Parameters: Specifically define boundaries of the research project, and 5. Narrow Parameters: Boundaries of the research project that can and have to be made more specific.

Following a "speed dating" motif, students are timed and only have two minutes to give peer feedback on an aspect of another student's research poster before the alarm sounds and they must move to the next poster. The "speed dating" lessons are conducted several times throughout the year, as students add elements of their studies to their research posters, and the process is recursive and ongoing. These assessments meet William's (2011) two requirements for being described as "formative," since they include instructional and curricular activities that result in improvement in performance, and the learner must act upon the evidence gathered to improve their learning, including seeking specific help from peers and reflecting on ways to move their learning forward.

AP Research Symposia

When discussing Michael's future AP Research lessons, he and the professor also created scaffolded learning opportunities to improve students' confidence in their ability to discuss and disseminate their research. The professor suggested an authentic performance assessment opportunity that resulted in an online Research Symposium, which first took place in April of 2021 via Zoom. The symposium was established for the high school students to practice presenting their research and receive actionable and timely feedback from graduate students, who were also practicing, certified teachers taking a graduate level course in research themselves.

The symposium was scheduled just prior to the students presenting their real Project Oral Defenses (PODs), so that they would view the symposium as an authentic dress rehearsal to practice what they would need to do for a score the following week, since the course culminates in an academic paper of 4,000–5,000 words (75%), accompanied by a performance, exhibit, or product where applicable, and a presentation with an oral defense (25%).

To avoid any potential College Board conflicts of interests, the professor did not observe the high school students or listen to the specific feedback on the papers and PODS, since she is trained as an AP Research Table Leader and Reader and sometimes participates in the AP Research Reading. Throughout the collaboration, Michael and professor carefully read all materials provided by the College Board to make sure that they were not violating any of their rules. Though the College Board prohibits adults from correcting student work, they are permitted to proofread, point out strengths, errors, and areas in need of improvement.

The research symposia have improved both the students' and teachers' confidence in presenting their scholarly work and even their overall success in the AP Research course and graduate research classes. The first virtual research symposium was so well-received by the high school students and early career teachers that they continued to hold them in November and April of 2021–2024 and are planning to continue this practice moving forward.

Reflections from the Research Symposia

When asked to reflect on the experience, one of the high school researchers who participated in April of 2024 wrote: "A lot of the feedback that I received had to do with defending the purpose of my statements. Especially when I offered quotes from websites and specific people, I was told to make sure I clarify who and what these websites and people are known for, providing credibility and preventing confusion. I plan to make sure I describe or briefly mention the purpose of me including each source. I also need to proofread my work to make sure I don't have any slip ups I might have missed." Another student posited, "I got the feedback that my presentation was overall good as I was passionate about it, however I could do a better job at defining some of the key terms used that wouldn't be general knowledge. I also was told that my slides could be more uniform in their organization (font, sizing, colors, and pictures) to make it look more professional. The last piece of feedback was to add subheadings under the pictures that I took

of Atlantic Beach, so the reader knows they were my own. I plan to first add a slide in the beginning of my presentation for definitions and key words. Then, I will go through my slides and make them uniform to give a cleaner feel to the presentation. Finally, I will go in and label all of the pictures that I took as my own and the beach that the study was based on." Every year that the symposium is held, the AP Research teacher points out that the symposium serves as a critical culminating experience for his students, whereby they get to see how ready they are for the POD, based on this authentic dress rehearsal and provides feedback to them that, when acted upon, improves their PODs.

Similarly, the Queens College teachers (graduate students) embrace the experience as assessment in the service of learning and always have positive comments about the symposium and how it supports their own learning, while concurrently doing so for the high school students. Most recently, in April 2024, a teacher commented, "One concern I had over the course of the week leading up to meeting the two students assigned to me was that I wasn't sure whether or not my commentary would be helpful or useful to them. Despite my initial insecurities about my own writing abilities, I realized that my role as a mentor was not about being a perfect writer, but about offering guidance and support. I shifted my focus to providing constructive feedback that I would appreciate receiving myself. This change in mindset gave me the confidence to meet with the two students assigned to me. Additionally, I didn't know what to expect in terms of topics, but each student's topic was different from the next, and it was honestly refreshing to see and read about. That part was exciting to me." Another wrote, "Overall, my general feedback for Amy and Violet (pseudonyms) was to reread their action research papers aloud in a mirror with a pen. In doing so, they would be able to check for grammatical errors and issues with organization in their writing. I also advised them to be specific when referring to their data and the reasoning behind the limitations within their research. With that being said, I wish more schools across New York had students engage in action research to encourage real-world problem solving and student engagement."

In addition to the creation of the bi-annual research symposia, several other new practices have been implemented, including ETS professional development opportunities for the teacher in becoming a Rater (scorer) for the AP Research Reading, the creation of an informal parental network to link novice student researchers with outside experts in the fields that they are researching (something

that is encouraged by the College Board), additional in-class activities that promote timely and actionable feedback, and additional formative assessments leading up to the culminating tasks of the POD and academic paper.

How the AP Research Course demonstrates Assessment as a Pillar of Pedagogy

The position taken in this chapter is that when assessment functions as a pillar of pedagogy both teachers and learners experience it as a seamless interaction with instruction and curriculum in ways that influence student learning positively. The AP Research course illustrates a few examples of assessment and its interdependence with curriculum and instruction and its support for student learning in the classroom as described below.

The AP Research Course and Its Success Criteria

Unlike most courses for which a standardized test of achievement is administered at the end of the teaching-learning cycle, the AP Research course is quite transparent about its criteria for success as delineated in its rubric. The use of the rubric is an illustration of assessment as a pillar of pedagogy since it informs the design and implementation of instructional and curricular activities to help students achieve the goals and objectives of the course. There is a wealth of literature on success criteria and its use in assessment and other pedagogical processes for the purpose of guiding learners in what to look for in their performance and progress toward the goals and objectives at the end of a teaching-learning cycle (e.g., Moss & Brookhart, 2012; Fisher & Frey, 2014; Tomlinson and Moon, 2013); Wiliam & Thompson (2007).

Scaffolded Activities of the AP Research Course

The scaffolded activities of the course included an annotated rubric as curricular content in which instruction among peers was intended to provide feedback in ways that made features of the rubric understandable for students who were experiencing difficulty with it. Inherent in the concept of scaffolding are characteristics of probing questions that teaching persons or capable peers use for eliciting evidence of students' understanding, the provision of feedback, and the fading of instructional support until students can demonstrate their learning without it. These characteristics of scaffolding underscore the synchronistic function of assessment, curriculum, and instruction with learning processes.

Studies of classroom teaching have reported that participation in scaffolding activities facilitate learning (Fisher & Frey, 2013; Palincsar & Herrenkohl, 2002; Wentzel & Brophy; 2014). Moreover, the quality and type of feedback, a characteristic of the scaffolding concept, has been shown to play an important role in student learning (Black and Wiliam, 1998; Hattie and Timberley, 2007; Furtak et al., (2016); & Johnson et al., (2019).

The Symposia of the AP Research Course

The AP Research Symposia used an authentic performance assessment as a culminating experience of the course so that students could practice presenting their research prior to doing so on their Project Oral Defense (POD) for grading as required by the College Board. Using performance assessment formatively is another example of assessment as a pillar of pedagogy because it provides opportunities for students to receive feedback from teachers about aspects of their work still in need of improvement. For learners, the feedback also makes visible how well they make progress toward the goals and objectives of the course and their readiness to transfer what they have learned in the course to another context- the AP Research Paper administered by the College Board and ETS. The research literature on practice is quite robust on its effectiveness in improving learning (e.g., Benassi, Overson and Hakala, 2014; Harmon& Marzano, 2015; and The National Academy of Sciences; 2018).

How the AP Research Example Connects to Handbook Principles

The preceding AP Research projects/assignments exemplify and explicitly connect to Principle 1 of the Principles for Assessment in the Service of Learning which states that "Assessment transparency assists teachers, learners, administrators, and parents." The transparency of the AP Research rubric and scaffolded AP Research activities enables the students to better understand the components of the AP Research program and assessment and what they need to do to be successful. Additionally, they enable the teacher to better understand the various components of the AP Research program and what they need to do to be successful and to help support their students' success. A less obvious yet powerful benefit is that the transparency of the scaffolded activities enables the school administration to better support the various components of the AP Research program and what their teachers and students need to do and lastly,

the transparency of scaffolded activities enables families to better support the components of the AP Research program and what their children need to do.

The preceding scaffolded activities also exemplify and explicitly connect to Principle 7 of the "Principles for Assessment in the Service of Learning" which states that, "Assessment quality and validity should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility." For learners, feedback from the formative assessments embedded in the scaffolded activities provides them with decisions and next steps to improve upon their AP Research Projects and PODs before being formally graded on them. Feedback from the formative assessments also provides teachers with decisions and next steps to improve their teaching and fill in learning gaps for students as they conclude their projects/PODs. Feedback from the teachers to administration regarding the success of the scaffolded activities, formative assessments, and PD opportunities enable administration to better support the research teachers with resources, financial support, coverage to attend ETS professional development, dissemination of information in the district, etc. Finally, feedback from the formative assessments and scaffolded activities enables families to better support their students because they have a better understanding of the expectations and how they can become resources for their own children and for other students in the AP Research program (i.e., serving as community "experts" on other students' projects, etc.).

Example Two: The Use of Performance Tasks in Mathematics Education

Historically, African Americans have faced systemic barriers to educational opportunities that have resulted in representation and achievement gaps over generations. The research literature reveals that these gaps can be explained in part by sociocultural differences in educators' and students' backgrounds and perpetuated by practices that reflect a lack of caring and empathy for the underserved from those who have the power to enacted changes (Gordon & Yowell, 1994). Additionally, the overall climate of academic institutions from K–12 to graduate schools, including their levels of inclusivity and support for diversity, can significantly affect African American students' engagement, sense of self-efficacy, and persistence in educational settings (Allen, 1992; DeFreitas, & Bravo, 2012). Negative experiences with faculty or peers can lead to feelings of isolation, instead

of a sense of belonging and nurturing in the school/learning environment (Cohen, 2022; McClinton, Mitchell, Carr, Melton, & Hughes, 2018; McGee and Bentley, 2017).

Community and home environments can also play a role in the persistence and performance of African American students in STEM courses, in general, and mathematics courses, in particular. Gordon (1996) speaks generally about the importance of familial support to the overall success of the teaching, learning, and assessment processes for learners. However, Kunjufu (2002) speaks about what is often a class and cultural disconnect between the expectations of middle-class teachers for parents to be "assistant teachers" at home and the capacity of parents to meet those expectations educationally, culturally, or financially. For example, depending on the socioeconomic status of the household, parents may not have the educational backgrounds (e.g., taken courses in trigonometry), cultural knowledge (e.g., knowledge of European history), or resources (e.g., money to hire a tutor) to supplement what goes on in the classroom. Additionally, parents or caregivers may have negative attitudes about certain subjects, like mathematics, because of negative experiences they had in school. They may speak negatively about their experiences in mathematics in social conversations or in the presence of their children. Children may, in turn, adopt many of those attitudes from their parents or caregivers. Kunjufu (2002) wrote:

> *I am very concerned with the math disparity between the races. Upper-grade African American youth are doing very poorly in math. There are many factors contributing to this dilemma, but in the chapter for parents, I implore you to never tell your children you were not good in math or you did not like it. Children begin to think their failure is genetically driven and it's not relevant. Parents must encourage their children to reach their full math potential* (p. 134).

Parents, caregivers, community members as well as teachers must be keenly aware of the effects that the type and tone of their comments and feedback have on young people's sense of self-efficacy–that is, a person's belief in their ability to succeed in a particular situation or area such as mathematics. It follows that students who experience positive personal perceptions of efficacy about learning mathematics content will tend to engage in more mathematics tasks, select more effective strategies for performance, expend more effort, and persist more when faced with difficulties, as compared to those students who do not feel efficacious in this area.

Two Performance Tasks in Mathematics Education

The precalculus course containing two performance tasks were housed in an introductory level mathematics course at a Historically Black College/University (HBCU) in the mid-Atlantic region of the United States. The course was designed to introduce students to the properties of a variety of functions, including rational, exponential, logarithmic and trigonometric functions–just to name a few. The focus of this discussion is on trigonometric functions–specifically sine and cosine functions. Professors had great latitude when it came to how they designed and taught the course; however, all students enrolled in the various sections of precalculus had to take a departmental final examination developed by a committee of mathematics professors.

Two examples are presented to demonstrate how a portfolio of assessment strategies, and most notably performance tasks, were used to increase African American students' engagement, motivation, effort, sense of self-efficacy, and learning in the precalculus course. Notably, Delpit (1995) wrote about the necessity for teachers to learn from and about their students. Therefore, on the first day of class during the semester, the precalculus professor asked students to share some information about their programs of study at the university and discuss frankly and freely their previous experiences in mathematics courses. Additionally, a survey was administered to the students to obtain anonymous responses about their attitudes about mathematics. The data and information from the discussions and survey were used to guide the instructional and assessment processes for the course. Specifically, the following course goals, learning objectives, and performance tasks were included on the precalculus course syllabus:

Course Goals: (1) To improve the mathematics learning experience of African American college students—that *is, to improve their attitudes about and performance in mathematics—especially among non-STEM majors* by implementing instructional and assessment strategies that are socio-culturally and socio-cognitively responsive to students enrolled in the course (proximal); and (2) to increase the representation of African Americans in STEM fields (distal).

Lesson Learning Objectives: 1. Students will be able to use the properties (e.g., amplitude and period) of trigonometric functions (e.g., y=a sin bx) to explain and represent quantitatively real-world situations. 2. Students will be able to explain/model how sine functions with different periods (radio waves, microwaves, x-rays,

gamma rays, etc.) relate to real-world phenomena; and 3. Students will be able to graph the average number of daylight hours per month for Washington, D.C. over the course of two years, using information from a credible data source, and communicate how the information is related to a real-world situation. Objectives two and three were assessed through the use of performance tasks.

The precalculus course was part of the mathematics curriculum that could be used to satisfy a general education requirement for a degree program. Therefore, there were a diverse group of students from a variety of academic programs enrolled in the precalculus course. For students who were pursuing degrees in the arts, precalculus was most likely the last mathematics course they enrolled in during their college years (i.e., non-STEM majors); whereas, for students who were pursuing degrees in the sciences, precalculus was most likely one of the first courses in which they enrolled (i.e., STEM majors). The non-STEM majors became the focus for helping to accomplish the distal course goal of increasing the representation of African Americans in STEM fields. To assist in accomplishing the distal goal—the evidence of which would not be apparent for decades in the future–the precalculus professor focused on accomplishing the proximal goal which was to create and maintain a high level of student engagement in the course, improve students' attitudes and senses of self-efficacy about their performances in mathematics, and help them achieve the learning objectives for the course. The steps in achieving the proximal goal were operationalized by incorporating students' assets, interests, and culture into the instructional activities and the assessment strategies and performance tasks of the precalculus course.

Fine Arts and Mathematics (Performance Task One)

The early foundational and very impactful research of Hale (1986), Irvine (1990), Ladson-Billings (1995a, 1995b), and others on culturally relevant and responsive pedagogy and its effectiveness with students of color was a welcomed reminder about how to maximize learning for all students. Therefore, when two students who were enrolled in a Fine Arts degree program asked if they could demonstrate their conceptual understanding about the properties of the graph of a sine function (e.g., the amplitude and period) by situating their performance in an Afrocentric cultural setting, the precalculus professor encouraged them to do so.

The students chose to write and perform a skit based on the 1990s television sitcom, *Martin*. The skit seemed like a perfect choice for two potential future screen writers or actors. In the skit, Gina explains to Martin the properties of the waves that a microwave oven uses to cook collard greens. Even Gina's choice to cook collards greens rather than kale had cultural overtones. Gina also explains how the magnitude of the period of the waves for cooking collard greens in a microwave oven differs from that of other kinds of waves with shorter or longer periods or wavelengths (e.g., radio waves, x-rays, and gamma rays).

The sitcom Martin was an award-winning comedy show that aired for five seasons in the early to mid-1990s. Cast members described Martin as a television show which had several features that appealed to young people: (a) young Black love, (b) a snapshot of the '90s African American culture and fashion, and (c) the genius of young African American comedians, like Martin Lawrence. Consequently, the students felt a connection to the show—both culturally and generationally. The Fine Arts students thought it was an excellent way to use skills they were learning in their degree program and infuse "fun" into their learning. They also recognized the need to meet with the precalculus professor several times before the performance of the skit to get feedback about how best to explain clearly and simply how "microwaves" are used (Shepard, 2000). A decision was made to explain how the period of microwaves differs from other types of waves, such as radio waves, which are used for broadcasting or the period of waves for gamma rays, which are used for killing cancer cells. Their research took them to a picture of the Electromagnetic Spectrum similar to the one depicted below.

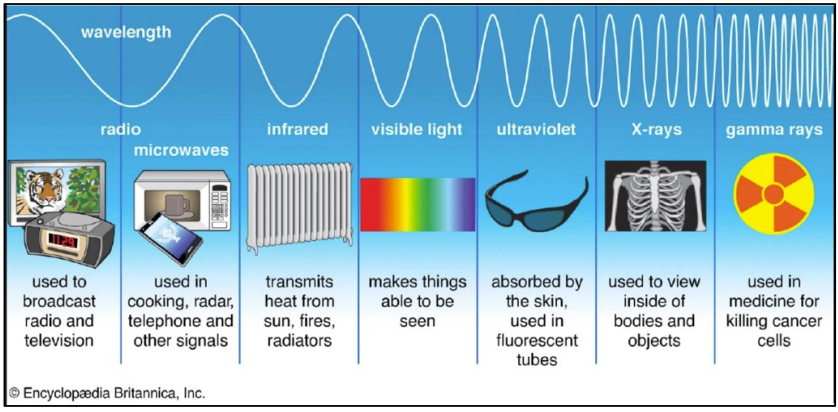


Figure 2

Collard greens are cooked in a microwave oven using waves with shorter wavelengths than radio waves which are used for broadcasting and longer wavelengths than gamma rays that are used in the medical field for killing cancer cells.

Additionally, a professor in the Mathematics Department was invited to listen and evaluate whether the students correctly explained the mathematical concepts about the period of a sine function and how it was used during their script. He gave the presentation "two thumbs up" and thanked the two students for "thinking outside the box" when it came to learning and demonstrating understanding of mathematical concepts. The mathematics professor also reminded the students that while the skit (performance task) was a creative way to demonstrate what they were learning during instruction; they would have to take a more traditional mathematics assessment–the two-hour departmental final examination–at the end of the semester. The mathematics professor encouraged the precalculus professor to prepare the students for this last assessment of the semester by administering two or three one-hour exams with items similar to the ones that appear on the departmental final examinations for precalculus.

Psychology and Mathematics (Performance Task Two)

All students enrolled in the precalculus course were assigned the following performance task: (1) consult an almanac or website for the average number of hours of sunlight for each month, January through December, of a particular year for Washington, D.C.; (2) plot the data on a graph for two consecutive years; and (3) write a few sentences about what they observed. Students discovered that their graphs of the data modeled two periods of the graph of a sine or cosine function. Some students were amazed that the average amount of sunlight per month for a particular city over the course of a year (or 2 years) can be modeled as a sine or cosine function.

There was a small group of students who were enrolled concurrently in precalculus and general psychology. The precalculus professor explained to them the relationship between the average number of sunlight hours per month over the course of the year and the psychological phenomenon known as Seasonal Affective Disorder (SAD). SAD is a mood disorder characterized by depression when there is less sunlight at certain times of the year. The students were encouraged to present their graphs from the mathematics performance task in their general psychology class so their peers could appreciate, in this example, a relationship between mathematics and psychology in a real-world context. The mathematics professor contacted the psychology professor and arranged for the students to present their graphs.

The members of the group created a PowerPoint presentation and shared the results of their performance task (mathematics) and their discussion about Seasonal Affective Disorder (psychology) with their peers and professors in their general psychology class. At the end of the presentation, the students received a round of applause from their peers and from their precalculus and psychology professors. They also received extra credit for their extra effort!

How the Precalculus Course Demonstrates Assessment as a Pillar of Pedagogy

As stated earlier in the introduction to this chapter, the concept of assessment as a pillar of pedagogy at the classroom level acts as a means for gathering information that can/should/is used to inform and document the interdependence of curriculum, instruction, and assessment in the service of and support for learning—for both students and their teachers. Thus, given that the focus in the precalculus

course was on trigonometric functions (e.g., sine and cosine functions)–it is reasonable to ask: "What assessment strategies were most effective for gathering information about students' readiness for learning the mathematics curriculum so they could successfully meet the goals and learning objectives of the precalculus course?" Similarly, it is equally important to identify and use assessment strategies that increase students' engagement, motivation, effort, attitudes, and sense of self-efficacy during instruction to facilitate learning. Finally, how does the use of these assessment strategies maximize students' abilities to demonstrate what they have learned after instruction?

The two performance tasks in the precalculus course, Fine Arts and Mathematics and Psychology and Mathematics, both demonstrate how the concept of assessment as a pillar of pedagogy was used in the service of learning in three ways: (1) knowing about learners and the curriculum through assessment prior to instruction; (2) allowing learners' choice in assessment during instruction; and (3) preparing learners for choice in assessment after instruction. A discussion of each illustration of the pillar concept in the mathematics performance tasks follows.

Knowing about Learners and the Curriculum through Assessment Prior to Instruction

Before the unit on trigonometric functions began, the instructor had in-class discussions with students about their backgrounds, interests, and motivations. Students talked about their major fields of study at the university, where they were from, and the reasons why they were enrolled in the Precalculus course. These discussions provided opportunities for the professor to learn about the students and for the students to get to know about each other. The professor also administered a survey to obtain students' responses about their attitudes about mathematics and previous experiences in mathematics courses.

Using assessment tools to generate information about learners' prior experiences in the content area in which they would be expected to demonstrate new learning, is a necessary first step of assessment as a pillar of pedagogy. The information obtained from these formal and informal assessments was used by the professor to design and organize subsequent teaching-learning experiences that were likely to elicit and sustain high engagement from the students in ways that would enable them to meet the goals and objectives of the precalculus course. A well-documented finding in the research literature on how learners learn is the relevance

of learners' prior knowledge in the construction of new learning (Bailey & Pransky, 2014; Schmidt & Marzano, 2015). Also well documented in the literature of formative assessment is the importance of using assessments to gather information of where learners are in a curriculum area of interest prior to instruction (Brookhart & McTighe, 2017; Hattie & Timperley, 2007; Wiliam, 2011).

Learners' Choice in Curriculum and Assessment During the Implementation Phase of the Lesson

During the implementation phase of a lesson, all the pillars of pedagogy (assessment, curriculum, and instruction) are in dynamic interactions with each other in support of student learning. For example, learners are expected to participate in curricular tasks and respond to instructional strategies that are used to help them learn. Assessments are also used to appraise how well they are doing and whether they are making progress toward the goals and objectives of the lesson. Of course, how well they are doing depends, in part, on whether students are motivated to engage in these pedagogical activities and remain engaged until task completion. In both performance tasks, Fine Arts and Mathematics and Psychology and Mathematics, the students were highly engaged in the curriculum and instructional activities. They were also receptive to the feedback given to them from results of informal assessments of those activities.

Task One: Fine Arts and Mathematics

The students' choice of the sitcom Martin as a setting for demonstrating their understanding of the mathematical concepts of the period and amplitude of sine and cosine functions was a way of letting their voices be heard. The Fine Arts students were familiar with putting on artistic performances thus, the use of authentic performance tasks was well-suited for them to demonstrate their learning in the mathematics classroom. The fact that they were allowed to use an authentic, socio-cultural, nontraditional way to demonstrate what they were learning is worth noting. It demonstrated that the instructor valued their preferred ways of knowing even if it was not traditional. One of the key outcomes for the course was to improve students' attitudes about mathematics and mathematics learning. The Martin skit by the Fine Arts students engaged the other students in the class and sent a message that they, too, could find common ground among their assets, interests, and mathematics.

Task Two: Psychology and Mathematics: The group of students who were enrolled in precalculus and general psychology also had a choice. They could choose to simply graph the average number of sunlight hours per month over the course of two years for Washington, D.C. and produce a sine or cosine function for the precalculus performance task or they could extend that mathematics performance task into an interdisciplinary project in which they had to apply and explain the real-world phenomenon of Seasonal Affective Disorder to a group of peers. The students chose the latter option and were rewarded with extra credit for their extra effort in their general psychology course. Furthermore, their choice helped them to appreciate the application of mathematics in real world situations.

While the three pillars of pedagogy (curriculum, instruction, and assessment) were in play in the two examples, what mattered most was the nature and quality of the responsiveness of students to these interdependent pedagogical processes during the implementation phase of the lesson. The high level of engagement, due to the professor's allowance of choice to students in curriculum tasks and choice in assessment to show what they know and can do, may have been the critical factors responsible for the learning that occurred during the implementation phase of the lesson. There is some support in the literature that giving students opportunities for active engagement in activities of a lesson facilitates their learning (Borich, 2014; Dean et al., 2012; Fisher, Frey & Lapp, 2011; & Peters & Kitsantas, 2009). Also, researchers in mathematics education have long recognized the importance of engaging students with the content and with each other about mathematics concepts to maximize mathematics learning during instruction (e.g., Mercer et. al., 2019; Resnick et al., 2010, Webb et. al., 2023). It is our contention that learning occurs when students focus their mental effort on learning activities involving the synchronistic and dynamic interactions of assessment, curriculum and instruction.

Preparing Learners for Limited Choices in Assessment after Instruction

At the end of the semester, after instruction, all students were required to take a two-hour final examination developed by a committee of mathematics professors. The departmental final examination was an assessment of students' learning. Furthermore, departmental final examinations tend to be more traditional, more standardized. The students do not have a choice.

To prepare students for the departmental final examination, it was important to familiarize them with the format and types of items that were likely to appear on the test. For example, some features of the format of the test were as follows: (1) it was a timed test; (2) it was administered on an individual basis–no collaboration with other students was allowed; and (3) some of the test items required students to show their work in order to receive partial credit. To facilitate and monitor students' familiarity with the format of the departmental final examination, the instructor administered two one-hour tests during the semester with content, item types, and formats similar to those that would likely appear on the departmental final examination. One purpose of the one-hour tests was to minimize levels of test anxiety among students so that they could maximize their performance. Assessment as a pillar of pedagogy works differently here when the attributes of the assessment become the focus of instruction. The professor conducted a test analysis on the results of each one-hour test to determine if there were any common areas of concern among the students' responses with respect to the content, item types, or formats. The results of each test analysis were used to provide corrective feedback to students–individually and collectively–where necessary. Empirical studies have underscored the effectiveness of the quality and type of feedback from assessment given to students to improve their learning (Hattie and Timberley (2007; Furtak et al., (2016) & Johnson et al., (2019). It is worth emphasizing, though, that students must act on the feedback in meaningful ways if it is to have an impact on their learning.

Application of the Principles for Assessment in the Service of Learning

Principle 1: *Assessment **transparency** provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.*

Students are better equipped to achieve course goals or learning objectives if they understand what the goals and objectives are, assume ownership of them, and can self-assess progress toward them (Nicol & MacFarlene-Dick, 2006). Understanding the goals and objectives means that there is significant agreement between teachers' expectations and students' conceptions of what is expected of them. One way to clarify expectations is to provide students with written statements of intended outcomes. Students enrolled in precalculus were given a course syllabus that outlined the use of authentic and culturally relevant and responsive performance tasks during instruction as well as a traditional departmental final examination which would be developed by a committee of mathematics professors at the end of the semester.

Principle 3: *Assessment design supports learners' **processes**, such as motivation, attention, engagement, effort, and metacognition.*

Culturally relevant and responsive pedagogy (Ladson-Billings, 1995a, 1995b) connects mathematics content to students' lived experiences and facilitates their engagement, motivation, attention, and effort in and understanding of the content. The performance of the skit, based on the sitcom Martin served as a means for teaching students about the mathematical concept of the period of a sine/cosine function as well as an end for assessing students' knowledge and understanding of the concept. The skit was an "instructment"–the combination of **instruction** and **assessment**. An "instructment" is more than a class activity. It also provides valuable feedback to students and teachers about how well students are learning the goals and objectives of the intended curriculum. The notion of an "instructment" builds on the work in the area of formative assessment (Black & Wiliam, 1998; Johnson, 1995).

Principle 5: ***Feedback**, adaptation, and other relevant instruction should be linked to assessment experiences.*

High quality feedback helps students improve their learning and performance. The quality of the feedback is, in part, a function of its relevance, target, timeliness, and tone. Hattie and Timperley (2007) note that feedback should be relevant to the task and address the needs of the student(s). Task-related feedback contains information about how well the goal of the task is understood, how well the task is being accomplished, and whether engagement and involvement in the task led to the attainment of the intended learning outcome(s). Feedback about performance on a task can be targeted at the individual or group level. While individual feedback to each student is ideal, teachers sometimes give feedback to a group of students about group performance. In this latter case, the usefulness of the feedback about the performance may be confounded by the perception that the feedback pertains to other members of the group and not to oneself. Students in the precalculus course received individual and group feedback before and after their performances. Finally, teachers should be aware of and sensitive to the tone–praise and criticism–when giving feedback. Practice the principle "Do no harm!"

Takeaways for Teachers

In this chapter, we make the claim that assessment as a pedagogical process of teaching is inseparable from other pedagogical processes of teaching (curriculum and instruction) when understanding and the improvement of learning are its primary focus. There are several takeaways from the preceding AP Research and Pre-Calculus examples that can serve as guidance for teachers who wish to use the metaphor of *Assessment as a Pillar of Pedagogy in the Service of Learning* in their classroom.

1. Knowing where students are in relation to new learning, where they need to be when a learning cycle ends, and what progress they are making toward where they need to be when a learning cycle ends all involve the interdependent relationships and functions of assessment, curriculum, and instruction.
2. Feedback from results of assessment about student learning is meaningful when it is used to fill learning gaps by teachers in the adaptations they make in curricular and instructional tasks for learning.
3. Students' mastery of learning goals and objectives depends, in part, on opportunities they are afforded for engaging in learning activities where assessments are linked to curricular and instructional tasks.
4. Assessments that offer students choice in modalities of representation of what they know and can do must be complemented by teachers' choice in modalities of representation of curricular and instructional tasks for learning.
5. Culturally relevant assessments for learning must be compatible with culturally relevant curricular and instructional tasks for learning.
6. A learner's engagement in assessment tasks is likely to be sustained when its results are used to design curricular and instructional tasks of interest to them and address their needs.
7. Assessments with embedded scaffolded instructional and curricular features are likely to influence student learning positively.
8. Features of assessment when used in instructional and curricular tasks are likely to influence students' learning positively.
9. Peer and teacher feedback from assessments influence students' subsequent engagement in instructional and curricular tasks.
10. Learning is facilitated when the interdependent relationships of assessment, curriculum and instruction are in alignment with learning goals and objectives.

Conclusion

When evidence is generated from assessments that students are learning and making progress in their learning, we can argue that assessment is in the service of learning. But a closer look at this claim reveals that for assessment to function in this way, it must be integrated with other pedagogical processes that share a similar purpose of assessment to inform and improve learning. As students engage in activities when these interdependent pillars of pedagogy are in operation, the evidence of learning and its improvement become visible when they demonstrate what they know and can do in a variety of ways (e.g., their verbal exchanges with peers and or the teaching person and digital, figural, graphical representations of their work). Analyses and interpretations of the evidence of the processes for learning can be used by the students and their teachers to inform changes in one or more pillars of pedagogy and learning processes that drive subsequent learning and teaching. In this chapter, examples from two disciplinary domains of AP Research and Mathematics Education were used to demonstrate assessment and its inseparability from other pillars of pedagogy that have learning as its central focus. Research support for the notion of the interdependence of assessment, curriculum and instruction and its relationship to learning is referenced, as are principles of transparency, assessment design and feedback, selected from the *Principles of Assessment in the Service of Learning.* Moving forward, the issues discussed in this chapter offer some guidance for a research and development initiative to generate empirical evidence of assessment as a pillar of pedagogy in the service of learning.

References

Allen, W. (1992). The color of success: African American college student outcomes at predominantly White and historically Black public colleges and universities. *Harvard Educational Review, 62*(1), 26–45.

Armour-Thomas & Gordon, E. W. (2013). *Toward an Understanding of Assessment as a Dynamic Component of Pedagogy.* An essay submitted to the Gordon Commission on the Future of Assessment in Education (Technical Report). Educational Testing Service.

Bailey, F., & Pransky, K. (2014). *Memory at work in the classroom: Strategies to help underachieving students*. Alexanderia, VA: ASCD.

Benassi, V. A., Overson, C. E., & Hakala, C. M. (2014). Applying science of learning in education. Washington, DC: Society for the Teaching of Psychology.

Black, P. (2018). Helping students to become capable learners. *European Journal of Education, 53,* 144- 159.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1)*,* 7–74.

Borich, G.D. (2014). *Effective teaching methods*: Research-based practice (8th ed). Boston, MA: Pearson.

Brookhart, S.M., & McTighe. J. (2017). *The formative assessment learning cycle* (quick reference guide). Association of Supervision and Curriculum Development.

Chatterji, M. (2012). Development and validation of indicators of teacher proficiency in diagnostic classroom assessment: A mixed-method study. *The International Journal of Educational and Psychological Assessment, 9*(2), 4–25.

Cohen, G. L. (2022). *Belonging: The science of creating connections and bridging divides*. New York, NY: W. W. Norton & Company, Inc.

College Board, (2024). AP Research Course and Exam Description. https://apcentral.collegeboard.org/courses/ap-research/exam

Dean, C., Hubbell, E., Pitler, H., & Stone, B. (2012). *Classroom instruction that works: Research-based strategies for increasing student achievement* (2nd ed.). Alexandria, VA: ASCD.

DeFreitas, S. C., & Bravo, A., Jr. (2012). The influence of involvement with faculty and mentoring on the self-efficacy and academic achievement of African American and Latino college students. *Journal of the Scholarship of Teaching and Learning, 12*(4), 1–11.

Delpit, L. (1995). *Other people's children: Cultural conflict in the classroom*. New York, New York: New Press.

Farenga, S., Joyce, B. A., & Ness, D. (2002). Reaching the zone of optimal learning: The alignment of curriculum, instruction, and assessment In R. Bybee (Ed.), *Learning science and the science of learning* (pp. 51–62). National Science Teacher Association Press.

Fisher, D., & Frey, N. (2013). *Better learning through structured teaching: A framework for the gradual release of responsibility* (2nd ed.). Alexandria, VA: ASCD.

Fisher, D., & Frey, N. (2014). *Checking for understanding. Formative assessment techniques for your classroom* (2nd ed.). Arlington, VA: ASCD.

Fisher, D., Frey, N., & Lapp, D. (2011). Focusing on the participation and engagement gap: A case study on closing the achievement gap. *Journal of Education for Students Placed at Risk, 16*(1), 56–64.

Furtak, E., Kiemer, K., Circi, R. K., Swanson, R., de Leon, V., Morrison, D. ,Heredia, S. C. (2016). Teachers' formative assessment abilities and their relationship to learning. Findings from a four-year intervention. *Instructional Science, 44,* 267–291.

Good, T., & Brophy, J. (2018). *Looking in classrooms* (10th ed.). Boston, MA: Allyn & Bacon.

Gordon Commission on the Future of Assessment in Education, (2013). *To assess, to teach, to learn. A vision for the future for assessment.* (Technical Report). Educational Testing Service.

Gordon, E. W. (1996). Toward an equitable system of educational assessment. *Journal of Negro Education, 64*(3), 1–13.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement Issues and Practice. 39*(3), 72–78.

Gordon, E. W., & Yowell, C. (1994). Cultural dissonance as a risk factor in the development of students. In R. J. Rossi (Ed.), *Schools and Students at Risk* (pp. 51–69). Teachers College: Columbia University.

Hale, J. E. (1986). *Black children: Their roots, culture, and learning styles (Revised edition).* The Johns Hopkins University Press: Baltimore.

Harmon, K., & Marzano, R. (2015). *Practicing skills, strategies and processes: Classroom techniques to help students develop proficiency.* West Palm Beach, FL: Learning Sciences International.

Hattie, J. A., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan, 89,* 140–145.

Hughes, G. B. (2010). Formative assessment practices that maximize learning for students at risk. In H. L. Andrade and G. J. Cizek (Eds.), *Handbook of Formative Assessment* (pp. 212–232). Routledge: New York.

Irvine, J. J. (1990). *Black students and school failure: Policies, practices, and prescriptions.* Praeger: New York.

Johnson, C. C., Sondergeld, T. A., & Walton, J. B (2019). A study of the implementation of formative assessment in three large urban districts. *American Educational Research Journal,* Vol. 56, No. 6, pp. 2408–2438.

Johnson, S. T. (1995). "Instructments"–a phrase coined by Dr. Sylvia T. Johnson at Howard University as Principal Investigator of the National Science Foundation Project, "Developing and Evaluating Performance Assessments in College and Pre-college Mathematics".

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37.

Kunjufu, J. (2002). *Black students. Middle- class teachers*. Chicago, IL: African American Images.

Ladson-Billings, G. (1995a, Summer). But that's just good teaching! The case for culturally relevant pedagogy. *Theory Into Practice, 34*(3), 159–165.

Ladson-Billings, G. (1995b, Autumn). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal, 32*(3), 465–491.

McClinton, J., Mitchell, D. S.B., Carr, T., Melton, M. A., & Hughes, G. B. (2018*). Mentoring at minority serving institutions: Theory, design, practice, and impact.* Charlotte, NC: Information Age Publishing.

McGee, E. O., & Bentley, L. (2017). The equity ethic: Black and Latinx college students reengineering their STEM careers toward justice. *American Journal of Education, 124*(1), 1–36.

Mercer, N., Hennessy, S., & Warwick, P. (2019). Dialogue, thinking together and digital technology in the classroom: Some educational implications of a continuing line of inquiry. *International Journal of Educational Research, 97,* 187–199.

Moss, C. M.& Brookhart, C. M. (2012). *Learning Targets: Helping students aim for understanding in today's lesson*. ASCD.

National Academy of Science (2018). *How people learn: Learners, contexts, and cultures*. Washington, DC: National Academy Press.

Nicol, D.J., & McFarlene-Dick, M. (2006). Formative assessment and self-regulated learning: A model and seven principles for good feedback practice. *Studies in Higher Education 31*(2), 199–218.

Palincsar, A. S., & Herrenkohl, L. (2002). Designing collaborative learning environments: *Theory into practice, 41*(1), 26–32.

Peters, E., & Kitsantas, A. (2009). Self- regulation of student epistemic thinking in science: The role of metacognitive prompts. *Educational Psychology, 30,* 27–52.

Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: Association of Supervision and Curriculum Development.

Resnick, L., Michaels, S., & O'Connor, M. C. (2010). How well-structured talk builds the mind. In D. D. Preiss & R. J. Sternberg (Eds.), *Innovations in Educational Psychology: Perspectives on learning, teaching, and human development* (pp. 163–194). Springer Publishing Company.

Sahadeo-Turner, T., & Marzano, R. (2015). *Processing new information: Classroom techniques to help students engage in content.* West Palm Beach, FL: Learning Sciences International.

Schmidt, R., & Marzano, R. (2015). *Recording and representing knowledge: Classroom techniques to help students accurately organize and summarize content*. West Palm Beach, FL: Learning Sciences International.

Shepard, L. A. (2021, Fall). Ambitious teaching and equitable assessment: A vision for prioritizing learning, not testing. *American Educator*, 28–38.

Stiggins, R., & Chappuis, J. (2012). *An introduction to student-involved assessment for learning* (6th ed.). Upper Saddle River, NJ: Pearson.

Tomlinson, C. A., & Moon, T. R. (2013). *Assessment and student success in a differentiated classroom.* Alexandria, VA: ASCD

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. University of Chicago Press.

Webb, N. M., Franke, M. L., Johnson, N. C., Ing. M., & Zimmerman, J. (2023). Learning through explaining and engaging with others' mathematical ideas. *Mathematical Thinking and Learning, 25*(4), 438–464.

Wentzel, K. R., & Brophy, J. (2014). *Motivating students to learn* (4th ed.). New York, NY: Routledge.

Wiliam, D. (2011). *Formative assessments: definitions and relationships. Division H Invited Session: Formative Assessment: International Perspective and applications.* Annual Meeting of the American Educational Research Association, April 2011: New Orleans, LA

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: what will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: shaping teaching and learning* (pp. 53–82). Mahwah, NJ: Lawrence Erlbaum Associates.

Reimagining State Assessments in Service of Teaching and Learning: Design Principles for Instructionally Relevant Assessments

Aneesha Badrinarayan

Abstract

This chapter presents a comprehensive framework for redesigning state assessment systems to better support instruction and student learning. Historically, large-scale assessments have prioritized easily quantifiable outcomes, often emphasizing surface-level skills over deeper, more meaningful learning. This approach has inadvertently shaped instructional practices, leading educators to focus on test preparation rather than fostering critical thinking and complex problem-solving skills. In response, this work proposes six key design principles to realign assessments with instructional goals and educational equity.

The proposed principles include: (1) Authenticity—creating assessments that reflect real-world disciplinary practices and engage students in meaningful tasks; (2) Curriculum-Anchoring—designing assessments that align with high-quality curricula to incentivize and support effective teaching practices; (3) Educative Design—using assessments to build educators' understanding of instructional quality and effective learning strategies; (4) Developmental and Asset-Oriented Approaches—recognizing students' progress along learning trajectories and focusing on their strengths; (5) Cultural Responsiveness and Universal Design—ensuring assessments are inclusive and reflective of diverse student backgrounds; and (6) Instructionally Useful Data—providing timely and actionable feedback to inform instructional decisions.

By implementing these principles, state assessments can become powerful tools for driving positive instructional change, supporting teacher development, and fostering equitable learning outcomes. This chapter calls for a paradigm shift in assessment design—one that balances rigor with relevance and centers the holistic development of every learner.

The decisions states and districts make regarding what their assessments look like and what kind of information they produce inevitably shape instruction. Since the No Child Left Behind Act of 2001 ushered in an era of testing-based accountability for schools, large-scale assessments selected and administered by states and districts have been governed by a set of design decisions that emphasize easily generated, easily compared scores—even when these assessments are somewhat superficial proxies for the rich performances state standards set for student learning. This makes sense if state assessments play a narrow and siloed role, focused on sending up a red flag around school performance and triggering a cascade of follow-up actions. While this might be consistent with how designers intend for assessments to be used, there is little question that this is inconsistent with how assessments are actually used in practice (Badrinarayan 2024).

For example, consider the following reflections from high school teachers and university professors as they discuss how ELA assessments are influencing instruction, drawn from a focus group conversation as part of national focus groups conducted to better understand the relationship between teaching, learning, and assessment (State Performance Assessment Learning Community, 2022).

Professor: We're finding that students are admitted to our [college] programs based on high school grades and test scores that show some degree of proficiency in reading and writing…but students come to us completely unable to make sense of complex texts or write about them in deep, generative, thoughtful ways. One student came to my office hours last week in tears because she was overwhelmed by the reading load in my syllabus, which is just a few chapters of a couple of books each week. I was in utter shock when she told me she had not read a full book in the last two years of high school.

High School Teacher: That's not surprising to me. Our interim assessment cadence is really fast, and I need to keep kids moving through and give them lots of practice with the skills that are being measured, like finding the main idea. We don't

have time to read long books, so we do a lot of excerpts, short stories, that kind of thing, about a lot of different topics. But students get practice with a lot of different texts, so I think that's probably helping them get better and better at the skills I'm trying to teach them.

Professor: It seems like they are getting good at school and taking tests—this isn't what [ELA] is supposed to be about. Being a good reader and writer is about being a good thinker, a good maker of meaning. A skill like "finding the main idea" isn't something you just practice or develop in a vacuum—by the time students leave high school, they should be able to deal with nuance and complexity of negotiating meaning across multiple, robust texts, grounded in content and topical learning.

Teacher: Well, I have no idea what topics or texts students will have to read on the interim or state assessments, so I can't go really deep on any one topic or text. Those tests really focus on students being able to read any text about any topic, so it's less about bringing knowledge to the text, and more about just being able to decode and make meaning of the words on the page—the other stuff is for science and social studies to teach. The texts on tests are pretty simple and my students struggle with them as is, so I think the goals you're describing are a little too ambitious for most learners. Maybe for our advanced students! But so few of our students are close to "mastery" in terms of proficiency levels, I can't even imagine what the kind of performance you're describing would be—super mastery? What's the achievement level?

Professor: But that's because they're not actually engaged in meaningful comprehension activities if they just jump from text to text and skill to skill—that's not how you learn how to read and make meaning.

We can unpack this example—and many more like this—to pinpoint at least three key ways that state assessments influence the behavior of educators in ways that impact instruction:

1. **They signal both <u>what</u> and <u>how</u> to teach.** In the example above, the K–12 teacher notes specific cues around instruction that they are taking from systemic assessment experiences, such as intentionally emphasizing interaction with multiple, diverse, short texts over deeper sustained engagement with longer pieces, and practicing skills like finding the main idea over knowledge-building and related meaning-making of texts. This is consistent with decades-old adage

that "what gets tested gets taught," but highlights that this doesn't only happen at the level of content domains or major disciplinary subdomains—what is on the test influences what materials educators select to use with students, how they structure and scope student interaction with those materials, and what they prioritize in terms of practice- and knowledge-building.

2. **Tests signal the bar—and expected range—for student performance.** Large-scale assessments are usually developed with the expectation that they are measuring the "floor" of expectations for student performance. Like the standards they are intended to assess, they represent what all students should know and be able to do, without placing any limitations on teaching and learning that goes *beyond* the expectations of standards in developmentally appropriate ways. However, in practice, standards and assessments are often interpreted as truly reflective of the full range of expected instruction and performance. In the example above, the K–12 educator is clearly using the achievement or proficiency level descriptors associated with the state assessment as the markers of the expected range of student performance. Even when they are receiving direct feedback from a postsecondary interest holder about the need for something beyond what the test is measuring, the teacher assumes that this must be out of bounds of even what high-level students who have achieved grade-level mastery should accomplish.
3. **They reinforce orientations toward students and their potential.** The college professor opens this dialogue by stating a mismatch between what they believe students are *capable* of—reading and writing in deeply thoughtful and generative ways—and what students have been *given opportunity to develop* through school. However, the K–12 teacher explains this mismatch away by stating beliefs about what students are actually capable of doing, as evidenced by the test. "My students can't do that" is perhaps one of the most common reactions we hear when showing teachers examples of high-quality tasks designed to surface what students know and can do related to disciplinary standards and durable skills. When asked what changes need to be made for students to be successful, teachers often suggest changing the task, rather than the instruction leading up to it, to better reflect what students can possibly be successful with. These examples indicate that external, large-scale assessments reinforce beliefs that teachers have about what students have the potential to accomplish, creating a feedback loop that serves to shape instruction.

It would be disingenuous to suggest that tests have this kind of influence in a vacuum—state assessments likely have this influence *because* of how they have been positioned in the broader educational system. For example, student performance on state assessments is the overwhelmingly dominant factor in determining school accountability, which has historically had implications not only for school funding, but also teacher evaluations, school restructuring, and labeling of schools as "good" or "bad" by local communities and external parties. This positions state assessments as a central focus for school and district activities, and more decisions that impact instruction (e.g., instructional materials selection, scope and sequence determination, pedagogical emphases, and decisions about which students get access to different curricular experiences) are made in ways that align with signals from state tests. Similarly, because state assessments have been used in ways that are particularly consequential for adults, the specific technical tradeoffs that are made with regard to current large-scale instruments (e.g., emphasizing reliability and efficient scoring, prioritizing standards coverage over depth) are elevated as more robust and at a higher bar for quality than other instruments that make different, but equally reasonable, tradeoffs. These assessments and their technical qualities, then, are held up as a gold standard for other assessment design efforts, reinforcing their position as highly influential.

Together, these kinds of factors have created a culture in educational systems that values external, seemingly 'objective' assessment instruments and the data they yield. Even in subject areas where state assessments are much less consequential, such as science, we see similar behavioral patterns in response to state assessments from many different interest holders. Leveraging our understanding of how a range of users use state assessment information, assessments can be designed such that state assessments have a net-positive impact on instruction. We can do so by designing assessments (See Table 1) that:

- Are grounded in defined instructional shifts, from current practice, that assessments should be designed to incentivize and drive.
- Are designed and communicated such that the most proximate logical way to "match" the state assessment in local practice (e.g., interim assessments, classroom assessment resources) mirror activities that reflect research on how students develop disciplinary knowledge and practice.

- Recognize that what happens in the classroom is not limited to interactions between teachers, students, and the content of instruction alone. What and how students learn is shaped by decisions made by educators and leaders throughout the system. Indeed, many of the instructionally relevant decisions that state assessments are most likely to influence lie outside day-to-day teacher–student interactions. By focusing on the most impactful ways state assessments influence instruction, states can ensure that assessments have a positive influence on instruction without extending into purposes that large-scale external assessments are simply not well suited to address.
- Provide teachers and leaders with information that offers a significant perceived value-add over other kinds of information they already receive through their classroom, school, and district instructional and assessment practices and resources.

Table 1

Instructionally Relevant Decisions Interest Holders Make Based Partially on Assessments

Interest holder	Examples of instructionally relevant decisions and actions
Students	• Reflecting on learning goals and progress (metacognition) • Codesigning learning experiences that are relevant and meaningful • Actively engaging in disciplinary inquiry • Revising work to meet shared goals and expectations • Providing feedback on learning experiences
Fairness with harmful bias managed	• Providing opportunities for deep, sustained, and compelling learning • Enacting culturally and linguistically responsive teaching and reteaching practices that deepen disciplinary skills and understanding while supporting the development of disciplinary identities • Leveraging students' current understanding and experience as a foundation within which to anchor new learning • Engaging students in learning that mirrors the authentic behaviors and conceptual development of the discipline • Learning more about student learning

Table 1 (continued)

Interest holder	Examples of instructionally relevant decisions and actions
School and district leadership	• Scheduling decisions that enrich curricular opportunities for all learners (e.g., inclusive learning for emerging multilingual learners rather than pulling out of class time for language remediation, providing sustained and coherent time for science and social studies in K–5) • Using observation protocols and educator coaching that reflect the major instructional shifts of the discipline • Implementing equitable grading policies that match how learning toward standards should occur (e.g., allowing grading frequency and approach to be consistent with how students develop and make progress visible in high-quality teaching and learning, focusing standards-based grading approaches on deeper learning targets rather than superficial coverage) • Adopting and implementing high-quality curriculum and instructional materials • Investing in systemwide curriculum-based professional learning • Establishing responsive course options and pathways that create opportunities for learners
State leadership	• Developing and adopting coherent instructional materials, assessment, and professional learning policies • Incentivizing the use of high-quality instructional materials (e.g., reimbursement for the purchase of these materials, support for professional learning for high-quality open educational resources) • Tailoring state-offered professional learning to support key instructional shifts

Source: Badrinarayan (2025).

Assessments that do so acknowledge that, whether they intend to or not, large-scale assessments do impact instruction—facing this reality head-on means that these same assessments can be designed to contribute to more effective teaching and learning in both direct and indirect ways, rather than operating under the assumption that the ways large-scale assessments influence instruction are just "side effects" of tests designed to address other priorities.

The following design principles for large-scale assessments seek to position assessments that operate across many different schools and districts in ways that are in service of learning as described by the overarching ***Principles for Assessment in the Service of Learning*** (this volume). In particular, these design principles for large-scale assessments emphasize elements of the following Principles as those deemed (1) most appropriate for large-scale assessments, and (2) those most critically impacted by large-scale assessment design and use in both positive and unproductive ways:

> **Principle 1:** Assessment ***transparency*** provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.
>
> **Principle 2:** Assessment ***focus*** is explicit and includes purposes, outcomes, progress indicators, and processes that can be transferred to other settings, situations, and conditions.
>
> **Principle 5:** ***Feedback***, adaptation, and other relevant instruction should be linked to assessment experiences.
>
> **Principles 7:** ***Assessment quality and validity*** should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility.

Design Principles for Instructionally Relevant Assessment Systems

Based on evidence from assessment system design and implementation, as well as lessons learned by the Learning Policy Institute and members of the State Performance Assessment Learning Community while working within, alongside, and across states, a set of design principles emerge that govern assessments intended to support teaching and learning (Badrinarayan, 2024; See Table 2.) These principles are designed to:

- build upon current conceptions of alignment to standards.
- focus on the most discerning features of assessment system design—that is, those features that are most likely to distinguish between systems that lead to positive shifts in instruction vs. those that have neutral or negative impact on teaching and learning, while allowing for a range of ways states could enact these principles.
- triangulate among the most important instructional shifts; the key users; and the specific, evidence-based behaviors we want to influence; and
- walk the line between aspirational and doable—it is unlikely that any state's current large-scale assessment program meets all of these design principles, but it is imminently conceivable that they could make different design decisions right now to bring their assessments into better alignment with instructionally impactful goals.

This chapter provides an overview of the design principles documented by Badrinarayan (2024), including rationale for the design principle and critical features states and other large-scale assessment system designers may consider including in their assessment system to enact the principle within their systems.

This chapter uses the term "state assessment systems" to refer to the collection of assessment instruments, supporting documents, and assessment-related services states directly provide to districts, schools, and teachers with an expectation that their use will influence or change student performance on summative measures of achievement or proficiency used by the state. In some states, this might be concentrated on a single instrument, such as the end-of-instruction assessment, or it may occur in a model that leverages multiple assessment opportunities throughout the course of an academic year or course of study (e.g., through-year assessments). In other states, this might encompass a combination of an on-demand statewide test with locally selected and/or used instruments (e.g., local performance assessments). The principles are intentionally focused on features of required state summative assessments rather than optional local activities, because practitioners and decision-makers have stressed that *without* important changes to the state's formal assessment system—the system that educators and local leaders feel accountable to, and measured by—local efforts cannot be as impactful they deserve to be (Darling-Hammond & Adamson 2014, Darling-Hammond 2017).

Importantly, while these design principles have been developed to be ambitious, but achievable for states, they are strong indicators for potential positive instructional impact for many different kinds of large-scale assessment systems, such as those used in districts, in school networks, and as part of national systems (e.g., Advanced Placement, International Baccalaureate).

Table 2

Design Principles for Instructionally Relevant Assessment Systems

Principle. Instructionally relevant assessment systems are intentionally designed to be…	Summary Statement
Authentic. Assessments should highlight and center the key concepts, modes of inquiry, and ways of learning in the discipline.	The assessment system should include authentic tasks representing ambitious examples of learning and performance in the discipline. These tasks should reflect sophisticated and complete performances, signal and support engagement with science and engineering practices (SEPs) and crosscutting concepts (CCCs), and center sensemaking around meaningful phenomena and problems. These tasks should also engage students in sensemaking in ways expected in science and reflect the most important instructional shifts we want to see. This may include both individual and collaborative work, cascades and bundles of SEPs and CCCs, student choice either among tasks or about how to engage within a task, and more.
Curriculum-Anchored. Assessments are designed such that high-quality curriculum better prepares students for success on the assessment, it incentivizes the adoption and use of high-quality curriculum materials, and it supports implementation of high-quality materials.	Assessments should signal, incentivize, and support the use of high-quality curricula that center active engagement with the disciplines in ways that operationalize evidence from the learning sciences about how disciplinary knowledge and practice are developed. This positions assessments to provide information particularly useful to instruction; to encourage the use of instructional materials and models focusing on deeper learning; and provide students with meaningful learning experiences. In some cases, assessments may be designed to reflect high-quality instructional materials (HQIM); in other cases, they may be designed to be coherent with HQIM, but focus on complementing existing curriculum (e.g., providing extended transfer opportunities, providing opportunities to better attend to broader issues).

Table 2 (continued)

Principle. Instructionally relevant assessment systems are intentionally designed to be…	Summary Statement
Educative. Assessments build educator and student understanding of and experience with high-quality teaching and learning in the discipline.	The assessment tasks, student data, and supports for interpretation—should build educator understanding of what high-quality disciplinary teaching and learning look like, what kinds of tasks can develop and evaluate that learning, and how to provide feedback in ways that support progress toward these goals. Assessments should attend carefully to the learning of teachers and students alike and are designed such that teachers also feel they learned something meaningful about their practice. What assessments signal, measure, and provide information about should directly speak to the actions and decisions we want students, educators, and leaders to make—and help them learn how to do so and why it is important. This may be accomplished by incorporating performance tasks into the instructional process; releasing items, tasks, and student work so that educators can see the kinds of tasks students are being asked to accomplish and what scores reflect; involving educators in designing and scoring tasks; providing task and student response annotations; providing concrete next steps to take, aligned to features of high-quality teaching and learning in science and based on student performance profiles; and making student experience data available to educators and leaders to contextualize performance.
Developmental and Asset-Oriented. Assessments recognize what students *do* know and *can* do, and surface progress relative to students' own performance and along appropriate learning progressions.	Assessments should focus on providing all students with opportunities to show what they know and can do relative to sophisticated disciplinary meaning-making. This includes emphasizing scoring and reporting that focuses on recognizing facets of student understanding and supports student growth over time. Assessments should also provide information about student performance along extended, multiyear learning progressions as well as expected learning progressions within a learning sequence (e.g., accounting for how modeling is expected to develop).

Table 2 (continued)

Principle. Instructionally relevant assessment systems are intentionally designed to be…	Summary Statement
Reflective of and Responsive to Learners. Assessments follow principles of universal design and cultural responsiveness to ensure each learner is supported in making their thinking visible.	Assessments should reflect students' cultural and linguistic experiences, employ multiple modalities for acquiring information and working through tasks, and include opportunities for students to demonstrate their learning in a variety of ways.
Useful for Informing Decisions That Impact Instruction. Assessments are designed to produce relevant information at appropriate times to support decision-making.	Assessment data must be made available at times when it can be used to positively impact instruction. In some cases, this might look like getting assessment data to users in a more timely fashion, particularly if the assessment design is intended to support changes in instruction for the given/current cohort of students. However, it should be noted that timing is not necessarily a discerning feature—more timely assessment results are only useful to instruction if the information is designed to be supportive of instructional decisions at those intervals (e.g., through-year assessment design). In other cases, assessments may be designed to primarily help teachers reflect on their own practice and plan for improving their instruction for their next cohort of students. In these cases, states may intentionally decide to slow down the process of returning scores to students in order to allow teachers to engage in rich and educative scoring experiences that can have a direct impact on instruction but result in scoring on a slower cadence than more automated processes might produce.

Source: Badrinarayan (2025)

Below, each design principle is further developed with rationale, critical features, and examples of systems as appropriate.

Principle 1: Authentic. Assessments should highlight and center the key concepts, modes of inquiry, and ways of learning in the discipline.

State assessment efforts can serve at least three important functions in teaching and learning systems, including:

1. Signaling what students should know and be able to do as a result of instruction, aligned to the state's standards, portraits of a graduate, and other visioning policies and documents.
2. Providing an example of the kinds of experiences students should be engaging in as part of ambitious teaching and learning practices.
3. Providing information about how students are progressing toward expectations in ways that students, families, teachers, and leaders can use to influence instructional practices in positive and productive ways.

All three functions require that students, families, teachers, and leaders can (1) trust that state assessments are measuring the knowledge, skills, and abilities most important to the discipline; (2) look to state-provided assessments as exemplars of what student performance should look like; and (3) easily use the information and examples provided by the state assessment to backward map to what teaching and learning should look like. For example, in science, this means if we want district-, school-, and classroom practice to cultivate authentic sensemaking and problem-solving with science and engineering practices, crosscutting concepts, and disciplinary ideas, we need state assessment systems to include authentic opportunities for students to do so. Similarly, if we want students to make meaning of texts and write to think and synthesize in English classes, we need assessments that provide students with opportunities to do so at the level of sophistication we expect in the classroom, incorporating processes of drafting and revision that aim for deep analysis. Without state assessments worth teaching to—that highlight and center the most important disciplinary activities, as authentically as possible—many students will simply never experience high-quality instruction, as the educators and leaders responsible for their learning get caught in never-ending cycles of trying to "game" the state test.

In practice, instructionally relevant assessments that prioritize authenticity include the following essential features:

1. **Include high-quality performance tasks that mirror real world disciplinary activities that require the targeted domain.** Assessments should leverage high-quality performance tasks that mirror (in developmentally appropriate ways) how the key ideas and practices of the discipline are leveraged in real world contexts as a component of state assessment systems. Doing so may involve (1) centering tasks on robust phenomena, problems, and scenarios that ask students to apply multiple disciplinary practices, crosscutting concepts, and conceptual ideas (potentially across disciplines); (2) collaborative investigations, modeling, and argumentation; (3) student-driven research, synthesis, and communication of findings; and (4) student thinking in response to phenomena, problems, and scenarios or contexts that contain authentic uncertainty, requiring students to use their knowledge and practice to address.
2. **Focus on the most-needed instructional shifts and higher-leverage aspects of standards and other related learning goals.** All large-scale assessments are an act of sampling. Federal law in the United States requires that states assess the depth and breadth of their state standards (U.S. Congress, 2015)—while this is intended to ensure that decisions about resource allocation and consequences for school systems are based on comprehensive information, efforts to meet this requirement often obscure a truth about state assessments: that they simply cannot measure everything students need to know and be able to do described by state standards. Current approaches to state test design almost universally value breadth and coverage, giving tests a façade of "measuring everything"—but in doing so, they are most often *not* measuring depth, robust and iterative meaning-making, deeper engagement with disciplinary practices, etc. Instead, assessment design could make a different tradeoff. In a world where no test is actually measuring everything, states and assessment developers should carefully consider *the highest-priority instructional shifts state assessments can support and ensure those same elements are prioritized in the assessment design.* This includes both major instructional shifts (e.g., toward integration of practices and discourse with content) as well as elements of the standards and related learning goals that are going to (a) have the biggest impact on instruction, and (b) be most critical for future learning, when prioritizing how learning goals are translated to assessments.

For example, one of the most widely acknowledged shifts needed in science teaching and learning across grade levels is more support and deeper student engagement in the science and engineering practices, particularly around interpreting and using evidence and being able to be critical consumers of information (National Research Council 2012, NGSS Lead States 2013, Badrinarayan 2024). Accordingly, states may design assessments that foreground science and engineering practices, emphasizing using science ideas as part of making sense of provided data and distinguishing between the validity of different claims and sources (National Assessment Governing Board 2024). In this example, states may choose to emphasize these elements over other types of performance, such as questions that focus on comprehensively surfacing conceptual understanding of a given science idea—in practice, this would likely look like having fewer score points associated with each disciplinary core idea, and more score points associated with integrated bundles of practices and core ideas, connected to more integrated performances. Similarly, across math, English language arts, and science, two major shifts researchers and practitioners emphasize is the need for students to (1) incorporate discourse and social meaning-making into learning activities, and (2) integrate the knowledge and skills across many different standards. Assessments could be designed to both signal these kinds of shifts: for example, large-scale on-demand assessments could develop items and tasks that emphasize measuring the construct that emerges at the intersection of multiple standards, even if this means defining new performance expectations for assessment that are derived from state standards statements. Large-scale assessments could also leverage emerging technologies to simulate discourse between students and interactive agents, or leverage collaborative and discourse-centered activities from the classroom in on-demand assessment activities (Shepard 2000).

Principle 2: Curriculum-Anchored. Assessments are connected to, and informed by, high-quality curriculum.

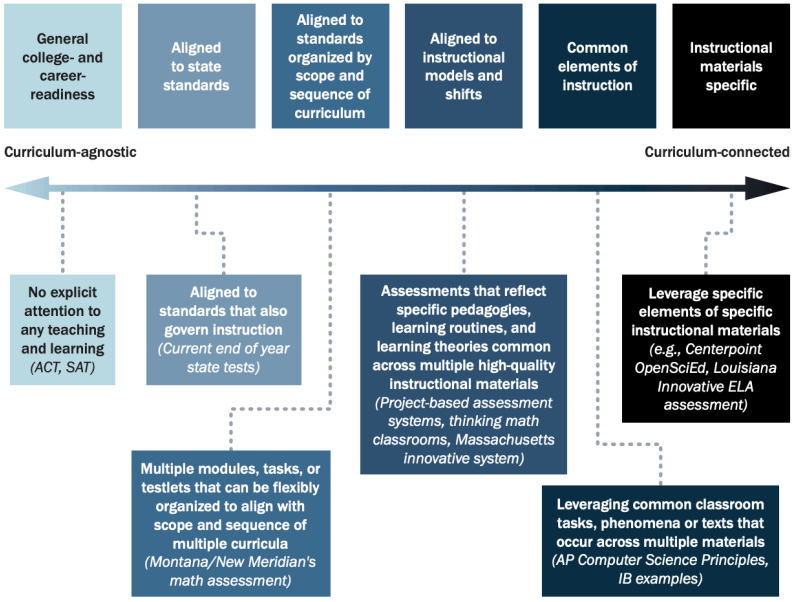
Traditional approaches to assessment design assume a unidirectional flow of influence in which state standards drive curriculum and assessment design independently, and, by virtue of alignment to standards, the assessments and curriculum are aligned. This "transitive property of alignment" approach makes some sense in theory, but it is not borne out by practice. Evidence from the learning sciences tells us that how students learn matters, and learning experiences influence how students make their thinking visible (National Academies of Sciences, Engineering, and Medicine, 2018). As a result, "curriculum agnostic" assessments that focus on end-of-instruction performance aligned to the letter of the standards, without attending to how students developed those ideas and practices, often result in superficial assessments that are not particularly useful to any high-quality curriculum context (Dadey & Badrinarayan 2022, Badrinarayan & Steiner 2023).

Instead, states should consider designing assessments that intentionally leverage features of high-quality curriculum (Dadey and Badrinarayan 2022, Badrinarayan and Steiner 2023). Across disciplines, leaders and developers have taken great care to ensure that high-quality instructional materials[1] (HQIM) and associated professional learning reflect the best evidence we have on how students learn within each of the disciplines, and do so in service of building mastery toward commonly established standards and related learning goals (e.g., EdReports n.d.). There are a range of approaches—some more tightly coupled to specific instructional materials, and others more loosely coupled to curriculum via common features of HQIM like theories of learning and instructional models—that states could explore to design "curriculum-anchored assessments" in service of better teaching and learning across their state systems (See Figure 1).

1 A growing body of research suggests that access to high-quality instructional materials, coupled with curriculum-based professional learning, is one of the most powerful factors in shifting teaching and learning practices. While decisions about curriculum and instructional materials adoption and implementation are generally left to districts and schools, an increasing number of states are committed to exploring ways to leverage state policies to incentivize the adoption and use of high-quality instructional materials. In this context, it is imperative that states carefully consider the relationship between instructional materials and state assessment systems

Figure 1
Ways to Anchor Assessment to Curriculum

Source: Adapted from Badrinarayan 2024.



By considering curriculum and assessment together, states can provide assessments that are more easily used by educators and leaders and ultimately have a stronger and more positive influence on curriculum implementation for students. Done well, these curriculum-anchored state assessments can measure knowledge, skills, and abilities in ways that attend to what and how students have had the opportunity to learn in science. These same assessments can also (a) incentivize the use of high-quality curriculum approaches and materials by making clear these approaches will help students be successful on state assessments, and (b) provide robust and ongoing support for connecting student progress to instructional decisions—ultimately acting as a tool to support local implementation of high-quality curriculum in ways that support every learner across diverse contexts.

It should be noted this design principle assumes state assessments should be more sensitive to instruction—in other words, performance on assessments provided by the state should vary based on whether students have received high-quality, impactful instruction, with students who have had better instruction performing better on state assessments, and students who have had more limited opportunity to learn performing less well. It has been well documented that increasingly external, large-scale assessments are generally less sensitive—and may be *in*sensitive—to instruction (Ruiz-Primo et al., 2012, D'Agostino et al., 2007). While we do not take issue with this finding, this design principle calls for us to challenge this moving forward—state assessments *should* reflect to some degree whether students have had opportunity to experience and reap the benefits of high-quality teaching and learning.

Instructionally relevant large-scale assessments intentionally account for curriculum by:

1. **Anchoring to curriculum in ways that reflect intended and likely use.**
   There are many different ways to make meaningful connections between high-quality curriculum and a state's assessment system (See Figure 1). The most appropriate connection within a given context will depend on the specific way a state intends for the assessment to influence instruction (e.g., incentivizing the uptake of HQIM or supporting the development of particular instructional practices connected to desired shifts) as well as historical and cultural contextual factors within the state (e.g., a history of strong state-provided support for instructional materials selection vs. a culture of teacher-led local curriculum development). States should consider both intended and likely uses to tailor the curriculum connection to maximize positive instructional shifts.

   When considering how best to connect with curriculum to support state goals, it will be important for states to consider how the assessment design will be both consistent with and complementary to high-quality curriculum. Striking the right balance between consistent and complementary features is essential to making sure the assessment has the desired impact on instruction. Features consistent with the curriculum should help educators and leaders clearly see the connections between the assessment and high-quality curriculum, and should support easy translation between information surfaced by the assessment and changes to instructional and programmatic practices and decisions. Features

that complement existing high-quality curriculum should provide a significant value-add to teachers beyond what they already have access to in using HQIM.

2. **Providing explicit guidance for how the assessment approach can be used with multiple HQIM available within the state**. No state mandates curriculum, nor is it likely that any single curriculum will be implemented in the same way over time. States should be clear about how their assessment system is compatible with multiple curricula while maintaining a strong stance on features of high-quality curriculum that the assessment signals and measures. State actions might include:

   - clearly communicating assumptions about learning and the underlying theory of learning governing assessment system design;
   - using curriculum surveys to identify the top three to five HQIM being used within a state and providing insight into the kinds of inferences that can be made about student performance relative to expected learning experiences;
   - building a library of assessment tasks that are assessing the same targeted standards but are intentionally designed for different HQIM and instructional models (e.g., project-based learning models, 5E instructional models, storyline curriculum models) for teachers to use based on their local curriculum decisions; and
   - coupling state-provided tasks with clear navigation routines that support educators making transitions from curricular contexts into and out of the assessment, to support coherence from a student perspective.

3. **Offering curriculum-anchored assessment professional learning.** One of the most powerful aspects of a curriculum-anchored assessment system is that it provides a clear avenue for coherent curriculum, professional learning, and assessment strategies. In curriculum-anchored assessment systems, the professional learning tied to assessments (e.g., designing, using, and scoring performance tasks) can become a form of curriculum-based professional learning, which evidence suggests is a particularly high-leverage strategy for improving classroom practice. States should elevate this aspect of the system, ensuring there are systemic opportunities for educators to develop an understanding of how to design and use assessments to support their specific

curriculum contexts. This aspect builds on evidence suggesting curriculum-based professional learning is particularly impactful on building instructional practice for teachers and meaningful learning for students (Short & Hirsch 2022).

Such a system might look like:

- connecting state-provided professional learning for task development, use, and student work analysis to curriculum (e.g., the instructional shifts governing many high-quality instructional materials, the instructional model within high-quality instructional materials);
- coordinating a vetted network of professional learning providers who offer specific curriculum-anchored professional learning connected to state-provided assessment resources;
- partnering directly with a specific technical assistance partner or partners to offer development and scoring workshops; and
- partnering with curriculum-related professional learning partners to sponsor and support professional learning communities to integrate assessment into curriculum-based professional learning.

Principle 3: Educative. Assessments build educator understanding of effective teaching and how students learn in the discipline.

A major concern about current assessment systems is they are used to justify ineffective, sometimes harmful, teaching and learning practices (NEA, 2021a, 2021b; AFT, 2024; Armstrong, 2013). In assessment systems designed to support better teaching and learning, assessments—including the tasks, student data, and released materials and supports—must be designed such that they (a) build educator understanding of what high-quality teaching, learning, assessment, and feedback cycles in science look like, and (b) are meaningful learning experiences for students unto themselves.

By designing assessments to be educative to educators and students alike, state assessments can be transformed from something done to teachers and students—an act often grounded in distrust of teachers and "gotcha" moments for students—into a tool used in service of growth, continuous improvement, and cultivating

teacher capacity and agency. Assessments designed to include elements that build teacher understanding will improve instructional experiences and contribute to curriculum equity by the very fact of their implementation. Moreover, state and local leaders can use educative assessments as the foundation for high-quality professional learning that bridges the assessment to curriculum implementation. For example, educators within districts often note assessment scores alone often drive short-term moves that do little to support student learning and retention (e.g., adding a unit on the scientific method to address perceived student deficiencies in science and engineering practice, repeatedly practicing reading decontextualized reading passages to practice skills like finding the main idea). However, when teachers are invited to examine tasks and student work (e.g., as part of assessment development or scoring workshops) that are illustrative (often annotated explicitly) of what teaching and performance in science or reading should look like, they find their discussions shift toward deeper understanding of state standards, needed shifts in practice, and concrete next steps.

This principle is a particularly important element of assessment systems that are seen as valuable to teachers and students, making it essential for instructionally relevant assessment systems to sustain and scale. Systems that are designed to be educative may:

1. **Provide transparent access to high-quality tasks and student work.**
   Perhaps the single most important element of educative assessments is that educators have transparent access to high-quality tasks and student work from the state assessment. It is nearly impossible to use assessments to improve practice based on numerical scores alone. When teachers and leaders can see what students were asked, and how they responded, they can (a) make more principled decisions about how to interpret and use reported data, (b) understand what it looks like for students to demonstrate targeted learning in context, (c) connect student performance back to their own teaching practice, and (d) notice nuances in student thinking and performance to help pinpoint needed support for individual and groups of learners.

2. Center and value tasks that are illustrative of high-quality teaching. Tasks used as part of state assessment systems should themselves reflect what high-quality disciplinary teaching looks like. For example, this reflection may include:

- providing guidance for appropriate customization of tasks (e.g., use of a more local example, use of more relevant texts or contexts)
- providing opportunities for independent and collaborative thinking,
- ensuring assessment tasks follow a coherent storyline from the student perspective,
- offering students ways to build on prior knowledge,
- offering students ways to engage in choice and decision-making within the task,
- providing students with opportunities to iterate and revise thinking as new information emerges,
- focusing on students using thinking and conceptual understanding over vocabulary use, and
- including rubrics designed such that they help educators and students understand a progression of learning as well as highlighting useful feedback to support student growth.

In curriculum-embedded performance tasks, assessments that illustrate high-quality teaching may look like providing comprehensive support for implementing the task. Tasks that do this may look quite similar to high-quality lessons and units, and include supports for classroom discussion, links to build teachers' understanding of both the subject matter and discipline-specific pedagogical knowledge, and support for high-leverage disciplinary teaching practices appropriate to the nature of the task.

In on-demand tasks that are released or otherwise made available for educators and leaders, assessments could provide detailed annotations that focus on associated teaching and learning practices (as opposed to only highlighting alignment) with supporting resources linked.

3. **Ensure tasks provide meaningful learning experiences for students.** When done well, state assessment systems can design assessments such that students experience the assessment as a meaningful learning experience in its own right. To do so, state assessments might:

   - prioritize authentic, compelling phenomena and problems that involve true uncertainty and clearly matter to an authentic stakeholder (rather than generic "scientist" or "students");
   - leverage tasks that can surface what students know and can do in service of figuring out something relevant and new to them;
   - provide opportunities for students to reflect on their own learning and progress and reflect on how this task helped build and monitor that learning;
   - include opportunities for student choice and decision-making within tasks (e.g., selection among a set of comparable tasks or choices within the task around data or reading selections, arguments to be made);
   - offer opportunities for students' own ideas to be an important element of engaging with the task (e.g., when making a recommendation for a design solution);
   - provide opportunities to transfer understanding in ways that further build students' schema within a particular science area; and
   - provide educators with guidance, including professional learning and routines for implementation that support maximizing the learning experience for students.

4. **Provide support materials that connect assessments to high-quality teaching and learning.** States should prioritize making supportive resources publicly available, including task annotations and related resources, student work samples, and guidance for instructional and programmatic next steps. This provision may include links to external resources about the phenomenon or problem such that teachers could use these resources as part of classroom activities, connections to high-quality instructional materials that are publicly available, or recommendations to high-leverage practices that could be used at the school or district level.

Involve educators in assessment development, evaluation, and scoring efforts. As discussed above, one of the most powerful elements of instructionally relevant assessment systems is the connection to high-quality teaching and learning via professional learning. Developing systematic ways to involve educators in state assessment design, evaluation, and scoring efforts—and designing those efforts such that they provide educators with opportunities for collaborative sensemaking about tasks, student work, and connections to policies and practices that influence instruction—is essential for assessment systems to realize their potential to support teaching and learning.

Principle 4: Developmental and Asset-Oriented. Assessments recognize what students do know and can do, and surface progress relative to students' own performance along appropriate learning progressions.

Educators often note a major concern with current state assessment efforts is that while they produce a great deal of data describing and labeling whether students have met grade-level standards, they do not provide useful information about what students *do* know and *can* do in ways that can be used to make better decisions about instructional practice, curriculum supports, professional learning processes, and more. These data-rich but information-poor systems often encourage and reinforce ranking, labeling, and shaming students, teachers, and schools rather than pointing to productive next steps in service of growth and acceleration. Not only does this approach to assessment provide limited support for teaching and learning, but it also contributes to static, deficit-oriented narratives about students' disciplinary identities (e.g., "Student A is bad at science," "I'm not a science person") rather than more accurate and dynamic views that some students have simply not yet mastered some disciplinary learning goals and need additional support and opportunities to do so (O'Donnell & Sireci 2021; Papay et al., 2011; Boaler, 2011; Carey, 2014).

If states instead focus on assessments that seek to identify more precisely what all students do know and can do, they can provide teachers, students, and families with assessment information that empowers them to take productive next steps to support growth. Research on how students learn makes it clear students learn by building on prior understanding and experience, rather than simply accumulating isolated pieces of knowledge and skill. Assessments designed to surface facets of student understanding along within-year and multiyear learning progressions for all

three dimensions of science standards can help students, families, educators, and leaders better identify assets of student thinking that can be leveraged as part of future learning opportunities.

Developmental and asset-oriented large-scale assessment systems may:

1. **Assess learning along learning progressions.** Instructionally relevant assessments should allow educators and students to identify facets of student thinking that educators can build upon. Because most states' standards are built upon intentional learning progressions that spiral across K–12, assessing along multiyear learning progressions can provide more accurate and asset-based accounts of what students know and can do, while still providing information that can be used to identify performance relative to grade-level expectations as required for state summative assessments under federal law.

2. **Use assessment tasks that provide all students with opportunities to feel successful by showcasing what they know and can do**. No student should leave an assessment feeling as though they could not demonstrate any understanding—this is both demoralizing for students and unhelpful to educators and leaders. Instructionally relevant state assessments should provide all students with opportunities to successfully make their thinking visible. This may be accomplished through a combination of on-demand items that range in complexity of sensemaking required as well as through compelling open-ended items and tasks that invite a range of thinking and response sophistication as students engage in sensemaking. Authentic, open-ended tasks in particular, when used in conjunction with asset-oriented student work evaluation and scoring processes, allow students to make more than simply a right or wrong answer visible. This strategy is essential for students who may not have had access to high-quality disciplinary instruction to still demonstrate assets in thinking that can be leveraged for next steps.

3. **Ensure assessments reflect how *learning* happens in each discipline, not just mastery of end-of-instruction expectations**. If assessments are to operate from an assumption that students who are not yet proficient have not had sufficient opportunity to learn, rather than that they are just somehow fundamentally incapable, we must assess from an assumption of learning, not mastery. Students do not develop mastery on specific end-of-instruction standards

by simply repeatedly "doing" the performance expected by the standard. For example, in science, students develop proficiency by using appropriate components of science disciplinary ideas, practices, and crosscutting concepts. In reading, students develop increasingly sophisticated text comprehension by iteratively connecting content knowledge and explicit reading skills. In all disciplines, students learn by being able to meaningfully engage with multiple representations of key content and practice, with appropriate scaffolds and opportunities for reflection and metacognition, over time and with increasing sophistication to support their use in service of meaning-making.

For assessments to surface student thinking along a wide range of proficiency levels, it is important that assessments reflect how students progress toward mastery within each discipline. This may mean assessments:

- move away from complexity frameworks that position recall as the simplest and/or least sophisticated performance;
- leverage features such as coherence from the student perspective, scaffolding/support, within tasks, and attention to learning contexts to better understand the degree of transfer expected by assessments to vary the complexity of assessment tasks;
- redefine proficiency levels (e.g., achievement-level descriptors and proficiency-level descriptors) to better attend to how learning happens in each discipline; and
- make more appropriate and realistic claims about the degree to which students can transfer and generalize their knowledge within disciplines at different stages of development (e.g., 3rd-graders are unlikely to have sufficient schema or lived experience to be able to demonstrate their learning in contexts that are extremely different than ones in which they experienced learning).

4. **Provide rubrics, scoring processes, and reports that highlight use of asset-oriented narratives of students that focus on facets of student understanding.** Instructionally relevant state assessment efforts should make available information about student progress that encourages growth-oriented feedback and next steps. Rubric and performance-level descriptors should be precise about what students know and can do and avoid highlighting gaps in student achievement as the primary way of describing performance.

Principles 5: Reflective of, and responsive to, learners. Assessments follow principles of universal design and cultural responsiveness to ensure each learner is supported in making their thinking visible.

All students have and are part of rich cultural traditions that govern how they learn, understand, and make thinking visible (Nasir et al., 2020, National Academies of Science, Engineering, and Medicine, 2018). Traditional approaches to state assessments that focus on standardization and "neutral" assessment contexts do not acknowledge students' interaction with content is dependent on lived experiences and how they developed their knowledge and skills—in other words, the sociocultural and linguistic structures surrounding the development of their disciplinary understanding. Thus, many current state assessments are not measuring what they intend to, inadvertently reporting on students' language and cultural familiarity rather than their science understanding and practice. While states have generally taken more care to attend to some features of universal design, such as using multiple modalities to present information about phenomena and problems, current state assessments do not often extend these features to other ways in which students engage with and respond to assessments, again limiting the degree to which the assessment is yielding trustworthy information about some students' learning (Randall et al., 2021, Randall et al., 2022).

Assessments that are reflective of and responsive to learners may:

1. **Assessments should act as windows and mirrors for students. Student performance on assessments are the result of student-task-context relationships**. No assessment task will have the exact same relationship to every student because students bring different experiences and assets to the table when interpreting and responding to a task. Instead, states and developers should seek to design assessment instruments that are in dynamic relationship with students—that is, some students will be "closer" to a given task while others will be farther, and the goal of assessment design is to vary this experience such that (a) all students feel seen and represented within an assessment, and (b) all students are supported in engagement with tasks, contexts, and/or ideas that might be less familiar to them. Designing assessment instruments that do this requires that tasks:

- Use a range of modalities (e.g., text, images, video, simulations) and types of text and information provided (e.g., graphs, charts, bulleted lists, different kinds of written sources).
- Center a range of perspectives and representation of who is a disciplinary knower, doer, and thinker (e.g., family or elder accounts, explanations that reflect different worldviews and experiences) as part of the information provided to students to shape tasks.
- Allow varied modalities and sensemaking routines to be valued in student responses. To the extent possible (e.g., open-ended tasks) and appropriate, assessments should strive for this to be true within a given task or item.
- Focus on assessments eliciting productive (not traumatizing) affective responses, rather than aiming for neutrality.
- Position assessment as learning opportunities—students should feel like they learned about someone's lived experience (related to their own community or someone else's) by engaging in the task.

2. **Use assessment tasks that connect to students' lived and learning experiences.** Assessments should provide opportunities to connect to student experiences through (a) the phenomenon/problem-based scenario and contextual information provided, (b) the items and student responses, and (c) reporting and feedback processes. It is essential these connections be as authentic as possible to avoid well-intentioned essentializing, stereotyping, and inappropriate assumptions about what is important to students and communities. This authenticity may look like:
   - leveraging known experiences from curriculum and high-quality instructional materials (connecting to Principle 2);
   - conducting student and community interest surveys and focus groups to determine relevant contexts and tasks;
   - providing choice to students among carefully designed tasks that measure the same construct but provide a range of contexts and ways to do so; and

- emphasizing tasks that leverage robust and common rubrics to allow students and educators flexibility around determining the specific elements of the task (e.g., students may be asked to design and justify a set of investigations to respond to a scientific question or area of inquiry but have flexibility in determining the exact nature of the experimental design, data to be collected, and implications for interpretation).

It should be noted "students" and "communities" are not static, monolithic entities—both communities and students are dynamic and ever evolving. This note suggests processes that seek to center student and community voice and experience should be approached as an ongoing dialogue rather than something undertaken once in time during an assessment development process.

3. **Position tasks to productively explore authentic and legitimate phenomena and problems that matter to specific communities.** While it may be challenging for states to generate tasks inherently interesting to each student engaged with an assessment, state assessments should strive to center tasks that are meaningful to a range of specific communities, such that the assessment as a whole is designed to be relevant to a true diversity of students, communities, and lived experiences. Tasks should position all communities represented as (a) more than stereotyped or victim-centered experiences, and (b) powerful doers and contributors to the discipline and the broader world.

4. **Centrally include tasks that authentically allow and encourage multiple perspectives and ideas as core elements of disciplinary reasoning and meaning-making.** Assessments should include tasks that (a) reflect a range of disciplinary ways of knowing (e.g., that there are multiple ways to "do science"); (b) position students' own ideas and perspectives—reflective of a range of individual and cultural identities—as an important element of disciplinary reasoning (e.g., when considering the trade-offs for possible design solutions); and (c) encourage and provide opportunities, connected to the disciplinary expectations, to critically consider what factors (historical, contemporary, social, cultural, environmental) have led to conditions and/or evidence being explored in a given task.

Principle 6: Useful for informing decisions that impact instruction. Assessments are designed to produce relevant information at appropriate times to support decision-making.

Current assessments are often criticized for how irrelevant they are to instructional decisions. For example, data on current state assessments often do not become available to students, teachers, and leaders until the next school year is well underway, making it challenging to use current state assessments to inform even longer-term instructional planning and decision-making. Moreover, when the information does become available, the nature of the information—numeric scores with little guidance for their interpretation—makes it difficult to connect to meaningful instructional shifts. Together, these features of current assessments surface perhaps the most urgent concern about the instructional relevance of state assessments: that they are simply not able to influence instruction.

Research and practical experience from states makes it clear that across stakeholder groups, people often make decisions that are most "proximate" to the data they have access to, in terms of both timing and content. States should intentionally consider this aspect of data use when designing instructionally relevant assessments, asking the question "How can we provide stakeholders with the right information at the right time to drive positive instructional shifts in the discipline?" It should be noted that attention to likely assessment use must be part of the assessment design and development process for instructionally relevant assessments. Put another way, it is not enough for assessments to simply make information about student progress or proficiency available, assuming that any decisions about what to do based the information is the responsibility of other parts of the education systems within which the data are used—instead, assessments intended to be instructionally relevant take seriously the impact of assessments (consequential validity) and design accordingly (Messick 1989, Iliescu and Greiff 2021).

Assessments that meaningfully attend to use may:

1. **Make specific and explicit claims about instructional impact, and design accordingly.** States often make claims about increasing the quality and utility of feedback for students and teachers when designing assessment systems—but fail to specify what would increase the quality and utility, who should be using it, and what decisions and actions those specific actors should be making. States and assessment developers should make clear, explicit, and discerning claims about what specific instructional impacts they want their assessment to produce. Consideration of claims should also include attention to unintended consequences of actual use, and how the assessment is designed to protect against those unintended consequences that might produce a negative impact on instruction.

2. **Make available data "closest" to the desired instructional shifts.** States should carefully consider the instructional shifts they want to drive and identify how to make compelling data available through the assessment system to support that shift. Doing so may look like:

    - *Making available different information on student reports,* such as redesigning score reports to include information about opportunity to learn to contextualize proficiency/achievement data; including anonymized/aggregated information on student interest, experiences, and opportunity to learn to accompany score reports; or including annotated samples of assessment tasks and student work to accompany performance/achievement level descriptors and reports of student performance.
    - *Presenting data in new ways,* such as exploring different reporting categories connected to instructional shifts, or including information about student highlights and competencies mastered rather than scores and performance levels alone.
    - *Focusing on a broader range of information* that goes beyond scores and sub-scores, such as sample tasks, student work, analysis of what students' response patterns say about how they are thinking and needed additional supports, and descriptive callouts about student performance on specific aspects of an assessment (e.g., the performance assessments vs. scenario-based item clusters).

3. **Design assessment systems such that their content and timing aligns with intended use, and designs for likely use.** While it may seem like more frequent information (e.g., via through-year assessments administered three or more times a year) will necessarily be more instructionally useful, this is not always the case. External assessments nearly always cause some disruption to teaching and learning, so it is imperative assessments designed to be administered more frequently are either (1) worth the disruption and potential costs to teaching and learning, or (2) designed to become a value-added component *of* instruction. Additionally, assessments administered more frequently have the potential to exacerbate the harms end-of-year testing already does in terms of reinforcing ineffective teaching and learning.(Goertz et al., 2010, Dadey et al., 2023, Datnow and Hubbard 2015) Indeed, when asked about current needs in assessments, local leaders frequently raise concerns about the quality and use of currently available interim assessments, which often derail well-designed and evidence-based instructional plans and models. It will be essential that states do not inadvertently position state assessments (e.g., through-year designs) to do the same.

As states consider the timing and cadence of assessments they use and make available, it will be especially important to consider how measuring different aspects of disciplinary standards influence how educators should make sense of resulting information. For example, science and engineering practices and crosscutting concepts in science standards are expected to build over the course of entire grades and grade-bands, while specific disciplinary ideas may be addressed in more modular ways. An assessment administered earlier in an instructional progression would have to be sensitive to these differences and help educators understand how to interpret the resulting student work.

Conclusion: Making a Different Set of Trade-offs

When state education departments view instructional impact as just as important as producing data to serve program monitoring, they can make different trade-offs in system design (See Figure 2), ultimately yielding better information and better educational experiences for young people.

By centering features of assessments that support better student learning experiences, teacher practice, and systematic supports and decision-making, we

can create assessment systems that have a "net positive" impact on instruction. The design principles detailed here reflect ambitious, but accomplishable goals for our assessment systems—and large-scale systems, including states as well as national and international programs, are already on the path to making this work a reality. As systems move forward, keeping "positive instructional impact" as the North Star and centering decisions on specific instructional shifts from the current state of teaching and learning that assessments should support can help system designers make the best decisions within their local contexts for better assessments.

Figure 2

	Perspective A: the primary focus of state assessment design should be on program monitoring. Positive instructional impact is nice to have, but not a must have.
Information prioritized	• Answering the question: are students proficient in the standards? • Coverage of as many standards as possible • Aggregate scores that can be used for comparing overall performance for schools, districts, and student subgroups at the macro level • Achievement/performance level descriptors relative to grade-level standards
Funding Allocation	• Assessment vendors and technical assistance partners who support test development, psychometrics, test maintenance, etc.
Scoring	• Prioritize machine scorable items. • Teacher involvement is a nice to have, not a must-have; teachers may be invited to participate in some content development and review activities, but teacher involvement is not central to the goals of the assessment. .
Student experience	Students experience the test as completely separate from instruction; see it as an external judgement of their abilities in the tested area.
Implications for test design	• Prioritize selected response items that can be developed and scored quickly and at low cost. • Prioritize coverage, focus on the of the easiest-to-assess content • Acceptable trade-off that teachers, students, families do not see the assessment items, and only see score reports

Figure 2 (continued)

	Perspective B: the primary focus of state assessment design should be on positive instructional impact. The assessment should also surface program monitoring information, but positive impact on instruction is a non-negotiable.
Information prioritized	• Answering the question: how well do students understand and use the standards? • Sufficient but not comprehensive coverage, emphasis on opportunities to measure depth, disciplinary practices, and deeper learning • Individual and aggregate scores as well as examples of tasks and student work that help educators and leaders make shifts that improve instruction • Achievement/performance level descriptors that more precisely pinpoint what students know and can do (e.g., along multiyear learning progressions).
Funding Allocation	• Assessment vendors and technical assistance partners • teachers, LEA, and SEA capacity • Substantial proportions of the resources are more directly supporting teaching and learning rather than only test-related activities.
Scoring	• Prioritize teacher-scored tasks alongside machine-scored items. • Teachers are an integral part of assessment development and scoring because participation in these activities builds educator understanding of standards and the kinds of teaching and learning that are needed to excel in the discipline (and on the assessment).
Student experience	Students experience the assessment as coherent with their instructional experiences, and as a meaningful experience in its own right. The assessment is engaging, motivating, and connected to what students have been asked to engage with in the classroom.
Implications for test design	• Include tasks that model instructional shifts and sophisticated performance expectations. • Prioritize surfacing information that signals and provides information about the most important instructional shifts, including disciplinary practices, disciplinary sense-making, and application of concepts and practices to relevant situations. • Teachers must have access to at least a subset of the assessment tasks to build practice and capacity.

References

American Federation of Teachers Assessment Task Force. (2024). *Real solutions for improving assessment.*

Armstrong, T. (2013). *15 reasons why standardized tests are problematic.* ASCD.

Badrinarayan, A. (in press). *Instructionally Relevant Assessments: What is the Role of Performance Assessments.* Learning Policy Institute.

Badrinarayan, A. (2024). *Design principles for instructionally relevant assessment systems.* Learning Policy Institute. https://doi.org/10.54300/111.528

Badrinarayan, A., & Steiner, D. S. (2023, June). *Positioning state assessment systems in service to teaching and learning: The role of high-quality curriculum in state assessment design.* Education First. https://www.education-first.com/wp-content/uploads/2023/06/positioning-state-assessment-systems-in-service-to-teaching-and-learning.pdf

Bennett, R. E., Darling-Hammond, L., & Badrinarayan, A. (Eds.). (2025). *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy.* Routledge.

Boaler, J. (2011). Changing students' lives through the de-tracking of urban mathematics classrooms. *Journal of Urban Mathematics Education, 4*(1), 7–14.

Carey, R. L. (2014). A cultural analysis of the achievement gap discourse: Challenging the language and labels used in the work of school reform. *Urban Education, 49*(4), 440–468.

Dadey, N., & Badrinarayan, A. (2022, April 21). *In search of the 'just right' connection between curriculum and assessment.* Center for Assessment. https://www.nciea.org/blog/in-search-of-the-just-right-connection-between-curriculum-and-assessment/

Dadey, N., Evans, C., & Lorie, W. (2023). *Through-year assessments: Ten key considerations.* National Center for the Improvement of Educational Assessment.

Darling-Hammond, L. (2017). *Developing and Measuring Higher Order Skills: Models for State Performance Assessment Systems.* Learning Policy Institute and Council of Chief State School Officers.

Darling-Hammond, L., & Adamson, F. (2014). *Beyond the bubble test: How performance assessments support 21st-century learning.* Teachers College Press.

Datnow, A., & Hubbard, L. (2015). Teachers' use of assessment data to inform instruction: Lessons from the past and prospects for the future. *Teachers College Record, 117*(4), 1–26. https://eric.ed.gov/?id=EJ1056748

EdReports. (n.d.). *Our process.* EdReports. https://edreports.org/process

Evans, C. M., & Taylor, C. S. (Eds.). (2025). *Culturally responsive assessment in classrooms and large-scale contexts: Theory, research, and practice.* Routledge.

Goertz, M. E., Oláh, L. N., & Riggan, M. (2010). *From testing to teaching: The use of interim assessments in classroom instruction (CPRE Research Rep. #RR-65).* Consortium for Policy Research in Education. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1023&context=cpre_researchreports

Nasir, N., Lee, C., Pea, R., & McKinney de Royston, M. (2020). *Handbook of the cultural foundations of learning.* https://doi.org/10.4324/9780203774977

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures.* The National Academies Press. https://doi.org/10.17226/24783

National Assessment Governing Board. (2023). *2028 NAEP science assessment framework.* https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/science/2028-naep-science-framework.pdf

National Education Association. (2021a). *Beyond the bubble: Americans want change on high stakes assessments.*

National Education Association. (2021b). *Principles for the future of assessment.*

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas.* The National Academies Press. https://doi.org/10.17226/13165

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states.* The National Academies Press. https://doi.org/10.17226/18290

O'Donnell, F., & Sireci, S. G. (2022). *Language matters: Teacher and parent perceptions of achievement labels from educational tests. Educational Assessment, 27*(1), 1–26. https://doi.org/10.1080/10627197.2021.2016388

Papay, J. P., Murnane, R. J., & Willett, J. B. (2011). *How performance information affects human-capital investment decisions: The impact of test-score labels on educational outcomes* [NBER Working Paper w17120]. National Bureau of Economic Research.

Randall, J., Poe, M., & Slomp, D. (2021). *Ain't ought to be in the dictionary: Getting to justice by dismantling anti-Black literacy assessment practices. Journal of Adolescent & Adult Literacy, 64*(5), 594–599.

Randall, J., Slomp, D., Poe, M., & Oliveri, M. (2022). *Disrupting White Supremacy in Assessment: Toward a Justice-Oriented, Anti-Racist Validity Framework. Educational Assessment, 27*(2), 170–178.

Conceptualizing and Evaluating Instructionally Useful Assessments

Scott F. Marion and Carla M. Evans

It seems silly to ask students to take a test that is not useful for improving their learning. However, many tests are designed to serve purposes other than instruction, such as monitoring and evaluating educational programs and school quality. Even still, far too many test vendors claim their tests are instructionally useful even if they were designed to serve other primary purposes. We rarely see evidence supporting these claims. In particular, we are concerned that teachers are often blamed for not using assessment results to improve their instruction. Doing so with assessments not designed to support instructional inferences is a Sisyphean task. Therefore, we wrote *Understanding Instructionally Useful Assessment* (Evans & Marion, 2024) to clarify our perspective on what it might take for an assessment to provide instructionally useful information.

Not surprisingly, our interest in defining and clarifying instructional usefulness parallels efforts to rethink balanced assessment systems more broadly. In the recent National Academy of Education publication, *Reimagining Balanced Assessment Systems* (Marion et al., 2024), the authors focus on rebalancing assessment systems to privilege rich classroom learning environments. The first part of the updated definition makes this point:

> *Balanced assessment systems and practices, as conceived by this volume's authors, are intentionally designed to provide feedback to students and information for teachers to support ambitious and equitable instructional and learning opportunities. This type of assessment system facilitates educator engagement in high-leverage professional practices such as quality formative assessment to support ambitious and equitable teaching* (Marion et al., 2024, p. 2).

As the definition shows, assessments that support learning and instruction are the focal point of a balanced assessment system. The authors emphasized that high-quality formative assessment practices can best support instructional uses. But what does it take for an assessment to be intentionally designed and implemented to support rich learning environments by providing instructionally useful information?

In the sections that follow, we first define instructionally useful assessments, then discuss key assessment design features that facilitate instructional usefulness, and conclude with some evidence requirements and areas for future research.

Instructional Usefulness Defined

> *"Although there is often a big difference between identifying a weakness and correcting it, identification can be a major part of the battle. By itself, a test score does little to identify the nature of a problem, only that there is one"* (Linn, 1983, p.179).

We started our journey into instructional usefulness by revisiting the wise words of some of the giants in our field. In addition to Robert Linn, assessment luminaries such as Eva Baker, Joan Herman, Peter Airasian, and many others contributed articles to a special issue in the *Journal of Educational Measurement* in 1983. However, even before 1983, assessment professionals have been concerned about having assessments support meaningful instructional actions.

We consolidated the ideas of many who came before us and defined an instructionally useful assessment as one that "…provides substantive insights about student learning strengths and needs relative to specific learning targets that can positively influence the interactions among the teacher, student, and the content" (Evans & Marion, 2024, p. 19). We further explored how instructionally useful assessments can support teachers by revealing insights through the assessment processes themselves, reporting results that shed light on student learning, or simply as a function of participating in the assessment (e.g., Agarwal et al., 2008).

Our definition, particularly the notion of substantive insights, follows from the conceptualization of the "assessment triangle" as described in the seminal volume, *Knowing What Students Know* (NRC, 2001). The authors emphasized that such insights are most likely to occur when there is coherence among the learning

goals and progressions, the assessments (observations), and the interpretative approaches. These substantive insights occur when the assessments support interpretations in light of the progressions by which students are expected to achieve domain competence.

We are aware that some might find this conception exclusionary. That was our intent. We have been concerned about claims that state standardized tests and commercial interim assessments can provide instructional insights, leading to considerable confusion about what it might take for an assessment to support instruction. We found Elmore's (2008) conception of the "instructional core" critical for framing our argument that instructionally useful assessment plays out in the interactions among students, teachers, and rich content within engaging learning environments.

Optimizing the conditions that make assessments more or less instructionally useful is necessary but insufficient for producing instructionally useful insights. We relied on Elmore (2008), Faxon-Mills et al. (2013), and Coburn and Turner (2011) to help us conceptualize how mediating variables influence instructional utility. In the case of instructional usefulness, the teacher's interpretive processes, decision-making, and instructional repertoires are critical mediating factors in influencing the instructional usefulness of assessments.

Beyond the skills of the teachers interpreting the assessment activities, other mediating factors include selecting an assessment that matches the enacted curriculum, having score reports that support productive actions, and ensuring that teachers have time and capacity to review and act on the assessment results. As we described in our book, the challenges of mediating factors increase as the assessment moves further away from the direct interaction between teachers and students. The presence of mediating factors is not an excuse for why an assessment does not yield instructionally viable insights. We highlight the importance of mediating factors so assessment designers and education leaders can understand the additional requirements and steps necessary for assessments to serve instructional purposes.

Design and Implementation Features

We spent considerable time thinking about the features of an assessment and the ways it is implemented and scored that would increase or decrease the likelihood that the assessment results could support instructionally useful interpretations and actions. We identified the following ten features as most relevant to explain why some assessments, and the information they produce, are useful for directly informing instruction, while others are less so.

1. Cognitive complexity and associated item types
2. Coherence with the enacted curriculum
3. Breadth of content standards and resulting grain size of results
4. Type of results (quantitative/qualitative)
5. Timing of results (e.g., ongoing, weekly, yearly)
6. Administration and scoring conditions
7. Allowable student responses
8. Student choice
9. Collaboration
10. Real-world and culturally relevant connections

Lorrie Shepard (2024) suggested that we organize the 10 features into three major categories. We provide additional details for these features, as categorized by Shepard, in the paragraphs that follow.

Representation of the Learning Goals

The ways in which the learning goals are conceived and represented are the most critical for influencing the potential instructional usefulness of an assessment and the associated results. We posit that coherence with the enacted curriculum, cognitive complexity of the test items, and breadth of content standards strongly influence the potential instructional usefulness of an assessment. This concept harkens back to Wiggins and McTighe's (2011) notion of backward design, where instructional designers are asked to begin with clear, desired learning outcomes and then design evidence of learning related to these intended outcomes. If an assessment is not embedded in the curriculum being taught, the teacher is

forced to undertake several additional interpretative steps to make sense of the curriculum-agnostic results in light of their curriculum.

As we explained above, deriving substantive insights into current states of student learning requires understanding students' location on a progression of learning rather than whether they have grasped it or not. Therefore, instructionally useful assessments must include a range of item types to probe different levels of cognitive complexity. Finally, instructional insights are generally tied to specific aspects of the curriculum and content standards. Therefore, assessment results should be at a small enough grain size so teachers can understand how to focus on specific knowledge and skills (Evans & Marion, 2024; Marion et al., 2024).

Just-in-Time Insights into Student Thinking

The timing and type of results can influence the instructional usefulness of assessments. At its simplest, results should be returned when it makes sense to offer "just-in-time" instructional adjustments so students can competently engage in the next curricular unit. Additionally, teachers need opportunities to "see" student thinking in order to derive substantive insights. Quantitative results can be challenging to interpret due to the obscure ways they are presented, especially for those without a deep quantitative background. We recognize the challenges associated with presenting actual student work for every response, but including at least some descriptive representations can help the interpretability of the full set of test results.

We also suggest that the degree of standardization required (or flexibility allowed) regarding the administration requirements, scoring conditions, and allowable student responses can affect instructional usefulness. These features are not as critical as the five features already discussed, but we believe these three features can influence the utility of the assessment results. Highly standardized administration and scoring requirements, as well as restricted types of allowable student responses, can limit students' ability to demonstrate their knowledge and skills. This might hinder teachers' understanding of what their students know and what they need to learn next to progress in their learning.

Sociocultural/Affective Aspects of Student Learning

We grounded our conceptualization of instructional usefulness in sociocultural learning theory and ambitious teaching (e.g., Ball & Forzani, 2009; Shepard, 2021). We are less certain about the degree to which student choice, collaboration, and real-world and culturally relevant connections influence instructional usefulness, but they are coherent with our theoretical orientation. Student choice could include flexibility in student learning goals, assessment targets, assessment methods, and flexibility in ways of demonstrating one's learning. Adults (and students) are expected to collaborate in cultural communities of practice, yet most of our assessments isolate students from these communities. Would teachers be able to derive more useful instructional insights if students were assessed in ways that more authentically allow them to show what they know and represent the world and their cultural communities? We suspect so.

Summary of Design and Implementation Features

The ten assessment design and implementation features all exist on a continuum. When most of the features, especially the first seven, are represented in ways to support substantive insights about student learning, the assessment is more likely to be instructionally useful. The last three features support instructional use because they support ambitious teaching practices and are often related to many other assessment features. For example, high-quality performance tasks can often support increased student choice, collaboration, and culturally relevant connections. Such tasks can generally yield descriptive results that provide insights into students' thinking.

It is not enough for only one of these features to be present. Many of the highest leverage features must be present because they mutually support one another. For example, assessment results that facilitate educators' insights into students' thinking are necessary, but insufficient to support instructional usefulness. If the assessment results do not provide substantive insights for teachers when they could feasibly adjust their instructional approach (e.g., timing is off) or teachers do not understand how the results relate to their enacted curriculum, the results will likely be of little instructional value.

While we can optimize features that make assessments more instructionally useful, assessments do not operate in a vacuum. For example, a high-stakes accountability use case would likely overshadow any potential instructional usefulness, even if the test was designed and implemented with many positive features (Evans & Marion, 2024; Marion et al., 2024).

Evidence of Instructional Usefulness

Claims are statements about what a product, person, or process will do under certain conditions. Claims are hypotheses that must be evaluated with evidence. Unfortunately, very little public evidence exists to support claims of instructional usefulness often made by test vendors who sell and education leaders who purchase tests (Diggs, 2019; Hill, 2020; Perie et al., 2009). What types of evidence would convince us that an assessment is instructionally useful? We describe four types of studies that may provide evidence to support or refute claims of instructional usefulness.

These studies are all somewhat post-hoc. That is, one needs the assessment and assessment items in order to conduct these studies. However, we argue that claims and evidence associated with instructional usefulness should start in the design phase. This idea is not new. It is at the heart of Evidence-Centered Design (Mislevy & Riconscente, 2006), which provides a formal framework for defining the claims about student knowledge, the observations (tasks) that would elicit evidence for those claims, and the interpretative models to evaluate the evidence.

- **Efficacy studies based on randomized controlled trials (RCTs)** are the "gold standard" of research; however, they are complicated to conduct in education. RCTs generally involve randomly selecting a sample of teachers from the population, randomly assigning a set of teachers to the "treatment" group (e.g., they get the assessment under question), randomly assigning another set of teachers to the "control" group, and then identifying a meaningful outcome variable and evaluate the difference between the two groups on this outcome. Despite the "gold standard" label, we believe RCTs are impractical, fail to account for considerable contextual influences, and overlook key insights from other approaches.

- **Cognitive laboratories**, also known as think-aloud protocols, ask participants to verbalize their thinking as they engage in an activity. We do this with students to understand how they engage with test items in the test-development stage. We can gain similar insights from teachers as they interact with student work and assessment score reports. In these studies, teachers are prompted to interpret student work or score reports as if they are talking to another teacher or if they are lesson planning. The interviewer probes for descriptions of specific interpretations the teacher generates from student work or score reports. The researcher also asks teachers to describe their specific instructional actions based on their interpretations.
- **Classroom observations** conducted by well-trained observers can shed light on how teachers make sense of assessment information. Compared with other methods, classroom observations have a major advantage: they provide insight into how teachers interpret and act on formative assessment activities and other informal assessments during instruction. Classroom observations also allow us to gather data about the effectiveness of teachers' actions based on their interpretations of assessment results and provide an important counterpoint to the types of interpretations and actions from summative or interim assessment activities.
- **Surveys** can provide information about instructional usefulness, but they require carefully crafted questions that pose scenarios to elicit evidence of how teachers interpret student work samples and score reports. Surveys that ask teachers if they found the assessment results instructionally useful are not worth the effort because of socially desirable responses. An advantage of surveys is that researchers can collect data from a representative sample of respondents, allowing for types of generalizations that are difficult to accomplish with smaller-scale studies, such as cognitive laboratories and observations. But, again, they have to be thoughtfully designed.

Evaluating Evidence

Once the data are collected, researchers should be able to indicate whether and how the assessment supported instructionally useful interpretations and actions. For example, teachers' interpretations and actions based on curriculum-embedded formative assessment opportunities could be compared to the interpretations and actions teachers derive from assessments further from instruction and the enacted

curriculum to evaluate claims about the potential instructional usefulness of a particular assessment.

However, teachers are not blank slates. They interpret new assessment information in light of what they already know about their students, including their learning strengths and needs in specific content areas. Therefore, evaluating evidence of instructional usefulness must be contextualized in terms of what teachers already know. If the additional assessment does not provide useful and usable insights, users and decision-makers must consider whether it is worth the time to administer an additional assessment if it only provides redundant insights.

Responsibility for Collecting Evidence

We argue that those making claims are responsible for providing the evidence to support them. Even less formal claims, such as those found on websites or other marketing materials, necessitate supporting evidence. We acknowledge that collecting evidence to evaluate the claims requires time and resources, but the types of studies we described above are not overly challenging to conduct.

While test vendors are primarily responsible for collecting evidence to support their claims, those making assessment decisions (e.g., district leaders) are also responsible. When shopping for assessments, district leaders and other decision-makers should ask for evidence of instructional usefulness before making a purchase. Evidence could come from the types of studies we suggested. Additionally, sample student work products or score reports resulting from an assessment's administration could allow district leaders to collaborate with their teachers to evaluate the extent to which the assessment results support instructional decisions and actions in their specific context.

Future Research

Collecting the types of evidence described above will help further our understanding of the features and characteristics that contribute to an assessment's instructional utility. However, additional research can help systematically examine various aspects of instructional usefulness.

We posited ten features that influence the likelihood of an assessment yielding instructionally useful information. We based our selection on what we could glean from the literature and our years of working closely with teachers to support their

assessment literacy. However, systematic research is necessary to provide insight into which features are more or less critical. Such research can shed light on the criticality of specific features (e.g., being able to examine student work) and the degree to which other features can compensate for shortcomings in other features.

Furthermore, rigorous research can help us understand how the various features interact with one another, as well as teacher expertise, school structure and culture, and other relevant factors. For example, we suspect that teachers with lower levels of assessment experience would be more likely to use assessment results if they are in schools with positive assessment cultures where they are provided time and support to interpret and act on results, compared to being in schools with weaker assessment cultures.

Finally, we need considerably more research into how score reports from standardized tests facilitate or hinder score interpretation and use. Most standardized assessments that teachers and students experience do not include descriptive or qualitative results in a way that easily supports inferences about student thinking. The quantitative results are generally presented in a formal score report. There has been extensive research into score reports (e.g., Hambleton & Zelinski, 2013), but not necessarily research into how score reports facilitate or hinder specific types of instructional inferences and actions beyond (re)grouping students.

Closing

Assessing students to gather information necessary for improving teaching and learning makes so much sense, it would be fair to question why we needed to write this chapter, let alone an entire book. Unfortunately, teachers and leaders are inundated with data, and there are many reasons why it is hard to translate the snowstorm of data into useful information. We need more research, but we hope we have helped conceptualize instructional usefulness in ways that motivate assessment designers and users to more seriously attend to what it takes to best support learning and instruction.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger III, H. L., & McDermott (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22, 7*, 861–876.

Ball, D. L., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education, 60, 5*, 497–511. https://doi.org/10.1177/0022487109348479

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement*, 9: 173–206.

Diggs, C. (2019, September 19). *Interim assessment? Didn't you mean formative assessment?* Center for Assessment CenterLine Blog. https://www.nciea.org/blog/interim-assessment-didnt-you-mean-formative-assessment/

Elmore, R. F. (2008). *Improving the Instructional Core.* https://achievethecore.org/content/upload/Improving%20The%20Instructional%20Core_Elmore%20Article.pdf

Evans, C. M., & Marion, S. F. (2024). *Understanding Instructionally Useful Assessment*. NY, NY: Routledge.

Faxon-Mills, S., Hamilton, L. S., Rudnick, M., & Stecher, B. M. (2013). *New Assessments, Better Instruction? Designing Assessment Systems to Promote Instructional Improvement.* Santa Monica, CA: RAND Corporation, 2013. https://www.rand.org/pubs/research_reports/RR354.html

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.). *APA handbook of testing and assessment in psychology, Vol. 3. Testing and assessment in school psychology and education* (pp. 479–494). American Psychological Association. https://doi.org/10.1037/14049-023

Hill, H. (2020, February 7). Does studying student test data really raise test scores? *Education Week*. https://www.edweek.org/leadership/opinion-does-studying-student-data-really-raise-test-scores/2020/02

Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20, 2*, 179–189.

Marion, S. F., Pellegrino, J. W., & Berman, A.I. (2024a). Reimagining balanced assessment systems: An introduction. In Marion, S. F., Pellegrino, J. W., & Berman, A.I. (Eds.), *Reimagining Balanced Assessment Systems*. Washington, DC: National Academy of Education.

Marion, S. F., Pellegrino, J. W., & Berman, A. I. (Eds.). (2024b). *Reimagining Balanced Assessment Systems*. Washington, DC: National Academy of Education.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Lawrence Erlbaum Associates.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment* (J. Pellegrino, R. Glaser, & N. Chudowsky, Eds.). National Academy Press.

Perie, M., Marion, S. F., & Gong, B. (2009). Moving towards a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28,* 3, 5–13. https://doi. org/10.1111/j.1745-3992.2009.00149.x

Shepard, L. A. (2021). Ambitious teaching and equitable assessment. *American Educator, 45,* 3.

Shepard, L. A. (2024, April 14). *Discussant comments. Is your test instructionally useful? How do you know?* [Symposium]. The Annual Meeting of the National Council of Measurement in Education. Philadelphia, PA.

Wiggins, G., & McTighe, J. (2011). *The Understanding by Design guide to creating high-quality units*. ASCD.

Practical Measurement for Improvement: Foundations, Design, Rigor

Paul G. LeMahieu and Paul Cobb

Educational systems today face persistent challenges that demand not only innovation but disciplined learning about what works, for whom, and under what conditions. Improvement science has emerged in response to this challenge, offering a structured, iterative, and evidence-based approach to addressing complex problems of practice. At the heart of this approach lies "practical measurement," a form of assessment that is embedded within the flow of professional practice and is designed to support real-time learning and continuous improvement.

This essay foregrounds three critical aspects of practical measurement in education: (1) the theoretical foundations of improvement science and its implications for measurement; (2) the design and implementation of practical measures; and (3) the technical quality and validity concerns that must be addressed to ensure responsible and equitable use. In contrast to discussions of abstract psychometrics and summative evaluations, this synthesis emphasizes how measurement can function as a tool that frontline educators, school leaders, and improvement teams can use to drive meaningful change.

By examining the purposes, essential attributes, uses, and technical criteria for practical measurement, this essay aims to articulate a coherent framework that supports its rigorous and responsible application in educational contexts. Importantly, this vision of measurement aligns with and reinforces other bodies of work that seek to reposition assessment as a tool for learning and advancing instruction, particularly the work on *Assessment in the Service of Learning* (AISL) championed by Edmund W. Gordon and colleagues (e.g., The Gordon Commission,

2013; Gordon & Rajagopalan, 2016; Gordon, 2020). AISL similarly foregrounds the role of assessment in directly improving teaching and learning, emphasizing formative, diagnostic, and student-centered approaches over primarily evaluative summative judgments. Practical measurement and AISL share a commitment to equity, actionable feedback, and a shift in power toward learners and front-line educators. Practical measurement, with its emphasis on real-time data use, co-design with practitioners, and improving teaching and learning represents a concrete instantiation of AISL principles in action.

Moreover, the knowledge developed through research on and experience with practical measurement, particularly how these measures can be integrated into supports for teachers and how validity-in-use can be attained, offers useful insights that can advance the goals and practice of AISL. The similarities in purposes for and contexts of these two assessment traditions lead to complementary similarities in both how we assess and how we can use the resulting data to guide improvement. Further strengthening the bridge between these traditions will enrich both and move the field toward more just and effective uses of assessment.

Theoretical Foundations of Practical Measurement for Improvement

Improvement science in education is grounded in six interlocking principles that reframe both the purpose and the practice of inquiry for improvement:

- Make the work problem-centered and user-focused,
- View variability in performance as the problem to solve,
- See the system that produces current outcomes,
- Use measurement to inform judgment and improvement,
- Apply disciplined inquiry, and
- Accelerate learning through networked collaboration (Bryk et al., 2015; LeMahieu et al., 2017).

Measurement, while appearing only in the fourth principle, is foundational to all six. Without evidence generated from practice, the identification of problems, diagnosis of causes, evaluation of interventions, and broader system learning would be impossible. Practical measures are the instruments that supply this evidence, not in abstract or delayed form, but in direct, timely, and actionable ways.

Practical measurement stands in contrast to traditional assessment systems that typically privilege purposes such as accountability or academic research (Solberg, L., Mosser, and MacDonald. 1997). These systems usually employ standardized assessments that are administered infrequently, measure only distal outcomes, and return results too late to inform ongoing improvements to practice. Moreover, these systems tend to be disconnected from the specific concerns and goals of practitioners. In improvement science, however, the goal is not only to track what is happening, but to understand "why," "how," and "for whom" change is occurring. Practical measurement is specifically designed for this purpose.

The centrality of practitioners in this process marks a significant epistemological shift. Rather than treating frontline educators as implementers of measurement-driven prescriptions, improvement science positions them as co-inquirers—designing, testing, and refining practices while contributing to shared knowledge. Measurement becomes a generative process, not just an evaluative one.

Design and Use of Practical Measures

Practical measures are not simply shorter or quicker versions of traditional assessments; they are conceptually distinct. They are designed to be:

- *Aligned* with a working theory of practice improvement that indicates what needs to be measured,
- *Relevant and meaningful* to those closest to the work and responsible for students' intellectual, social, and moral development,
- *Actionable*, informing specific decisions about changes that are likely to be improvements,
- *Minimally burdensome*, fitting within educators' existing workflows, and
- *Timely*, both in frequency of administration and in providing needed feedback (Takahashi et al., 2022).

These criteria enable practical measures to support disciplined inquiry cycles such as Plan-Do-Study-Act (PDSA), providing real-time feedback on whether a change is leads to improvement.

Clarifying Purpose: Improvement vs. Accountability vs. Research

A central premise of practical measurement is that the "purpose" of an assessment should dictate its appropriate design, implementation, and use. Measurement for improvement differs in crucial ways from both accountability-driven testing and from measures developed for use in traditional academic research.

Accountability measures are often summative, standardized, and high stakes. They are typically externally mandated and designed to evaluate whether schools, teachers, or students have met predetermined benchmarks. Accountability testing happens infrequently and is usually extremely time-consuming as it has to cover broad performance domains. The lengthy timeframes for processing and returning results make accountability assessments lagging indicators that offer little guidance for real-time adaptation of practice. They reinforce what the system currently values most.

Research measures are usually designed for internal validity, generalizability, and theory testing. They may assess constructs of interest to researchers that are not directly actionable or even relevant to practitioners' concerns. They are comprehensive and measure all aspects of constructs that might be relevant to the theory under examination.

Practical measures for improvement, in contrast, are designed to be integrated into educators' daily work. They capture leading indicators and thus enable practitioners to determine whether a change they have made is an improvement, and what adjustments they might need to make. Their primary goal is to support practitioners' ongoing learning from and in practice, not to determine whether benchmarks have been attained or to investigate relationships between theoretical constructs.

Design Considerations

Designing effective practical measures involves balancing technical rigor with usability. This includes:

- Anchoring items to specific factors relevant to the theory of improvement and the changes being tested–be they the processes or tools being changed or their intended and unintended outcomes,
- Ensuring that measures are sensitive to changes in practice,
- Structuring data collection and interpretation so that they are integrated into existing routines, (existing routines, and)
- Designing reporting formats to facilitate interpretation and action.

Iterative refinement of measures is essential. Initial versions of a measure may not capture important differences in practice, may be misunderstood by users, or may produce data too late to be helpful to practitioners. Developers must remain attentive to feedback, patterns of use, and the interpretability of results.

Myths and Misconceptions

As interest in improvement science has grown, so too have misunderstandings about the role and nature of practical measurement. Four common myths are especially important to address:

Myth 1: Practical Measures Are "Quick and Dirty"

The descriptor "practical" can misleadingly suggest casual or imprecise design. In fact, the opposite is true. Effective practical measurement requires high levels of rigor, creativity, and iterative testing. These measures must be simultaneously predictive of meaningful outcomes and usable within the real constraints of classroom, school, and district work processes. Far from being quick and dirty, they are often the result of assiduous development processes involving repeated design-test-revise cycles.

Myth 2: Researchers Design, Practitioners Use
The traditional division of labor between knowledge producers and users undermines the development of useful and meaningful tools to support ongoing learning and development. Improvement science insists on collaborative co-design: practitioners, researchers, content experts, and improvement specialists working together to determine what needs to be measured, design instruments, and test usability. This not only enhances technical quality but ensures that measures are interpreted and acted on in ways that respect the complexity of practice.

Myth 3: A Single Measure Can Serve Both Accountability and Improvement
Measures designed for accountability often distort the learning processes they aim to monitor. The high stakes of many assessments can encourage superficial compliance, teaching to the test, or gaming. Conversely, measures for improvement require space for safe exploration, including failure and iterative refinement. Attempting to use the same measure for both purposes compromises each. It is therefore essential that systems create dual infrastructures: one for accountability and one for improvement (LeMahieu and Wallace 1986; LeMahieu and Reilly, 2004).

Myth 4: Any Data is Better Than No Data
While data are essential, poor-quality or misaligned data can do more harm than good. Misleading indicators can obscure problems, foster false confidence, or direct attention away from real causes. Practical measurement emphasizes the need for the "right" data, aligned with the improvement aims, interpretable by practitioners, and capable of guiding productive action.

Validity Considerations

Although practical measures are situated in real-world settings, they must still meet standards of technical quality. Two complementary concepts: "*validity-for-use*" and "*validity-in-use*" are essential for understanding their trustworthiness (See, for example, Messick, 1989; Shepherd, 1997; Moss, Girard, and Haniford, 2006; Bond 2013; Smith, 2025).

Validity-for-Use

This refers to whether a measure accurately captures what it is intended to assess. It can be conceived and examined in a number of ways:

- Face Validity: Does the measure appear appropriate to users?
- Content Validity: Does it cover the relevant domain?
- Construct Validity: Does it assess elements of the theory of improvement that are the focus of specific improvement efforts?
- Predictive Validity: Does it correlate with future outcomes?
- Concurrent Validity: Does it align with other established measures?

Importantly, because practical measures are integrated into current work routines, their design must be especially sensitive to trade-offs between brevity and construct coverage. Strategies such as mapping onto existing research findings, cognitive interviews with users, and triangulation with other data sources can help build a robust validity argument. Validity-for-use is typically conceived of as a characteristic of the measure itself.

Validity-in-Use

Validity-in-use goes beyond psychometric properties of the instrument itself to consider *how* practitioners might interpret the resulting data and *what* actions those data prompt. A technically valid instrument can be misused if it does not produce data that can inform and support constructive action or if users lack the support, shared understanding, or conducive context needed to use it productively. This compels a co-creative process that attends to systems of use as well as to the measure itself. This precludes, for example:

- Measures interpreted in deficit-based ways that may exacerbate inequities;
- Data that are used for evaluation rather than improvement can provoke anxiety and surface-level compliance; and
- In the absence of routines for collaborative sensemaking, even good data may fail to lead to improvement.

Validity-in-use requires that its actual use be aligned with its intended purpose, and that its use is supported by adequate infrastructure, professional learning, and leadership. Validity-in-use focuses not so much on the characteristics of the measure itself but on how the measure is used and in the contexts in which it is used.

Implications and Challenges

In our experience, the development and use of practical measurements for improvement raise several issues and challenges that must be addressed if the measures are to be used appropriately and most beneficially to inform improvements of practice. These include (LeMahieu and Cobb, 2025):

Equity and Inclusion

Practical measurement holds particular promise for advancing equity. The enduring questions of improvement science: "what works, for whom, and under what conditions" compel attention to variability in performance and thus to extant inequities, thereby highlighting where changes are (or are not) benefiting historically underserved populations. However, this potential will be realized only if measures are co-developed with attention to diverse contexts and interpreted in ways that avoid deficit framing. Measurement must support—not obscure—efforts to reduce disparities in opportunity and outcomes.

Capacity Building

Using practical measures productively requires new capabilities across role groups. Teachers need support in interpreting data as they enact cycles of improvement. Coaches, facilitators, and leaders must cultivate collaborative data use practices. Researchers must develop new competencies in designing measures that are psychometrically sound *and* practically usable.

Sustainability and Scaling

Practical measurement is not a one-size-fits-all endeavor. What works in one setting may not translate easily to another. Nonetheless, by building libraries of field-tested measures, open-access repositories, and adaptable tools, resources can be created that support local adaptation while retaining shared learning. Networks for improvement can accelerate this process by testing and refining tools across a range of contexts, not with the fidelity of standardization but with integrity in adaptation as the focus.

Participation in Development

The essential attributes of practical measurement for improvement compel new thinking about how best to ensure that they fully realize their promise. Traditional thinking about assessment would place primary responsibility for (and therefore agency in) developing measures with the technical experts of the psychometric community. In doing so, a number of difficulties can arise. This is, in part, because traditional procedures for determining and assuring quality in assessments can constrain the form and focus of assessment. This too often provides evidence that teachers do not find useful. Practical measures must adopt forms and formats of assessment that provide evidence and analytics that practitioners find relevant, meaningful, and actionable. In our experience, this is most effectively accomplished when practitioners are centrally involved in development efforts, with those providing technical expertise included as members of the development team.

Conclusion

Practical measurement for improvement represents a fundamental rethinking of the role of data in educational improvement. It is not merely a technical tool, but a social and organizational practice—an engine for professional learning, informed judgment, and improvement. Properly conceived and skillfully implemented, practical measures can help educators see problems more clearly, test ideas more effectively, and work toward equity with greater efficacy and efficiency.

However, realizing this vision requires ongoing attention to design quality, contextual appropriateness, and the social dynamics of data interpretation and use. It demands that educators, researchers, and leaders reimagine their roles, not as isolated actors, but as partners in inquiry. By embedding rigor into relevance, and structure into responsiveness, practical measurement helps fulfill the core promise of improvement science: that we can, in fact, "*get better at getting better.*"

References

Bond, L. (2013). Toward a measurement science capable of informing and improving teaching and learning. In *Gordon Commission on the Future of Assessment in Education, To assess, to teach, to learn: A vision for the future of assessment (Technical Report).* Educational Testing Service.

Bryk, A. S., Gómez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better.* Harvard Education Press.

Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment (Technical Report).* Educational Testing Service.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice, 39*(3), 72–78.

Gordon, E. W., & Rajagopalan, K. (2016). New approaches to assessment that move in the right direction. In E. W. Gordon (Ed.), *The testing and learning revolution: The future of assessment in education* (pp. 107–146). Palgrave Macmillan.

LeMahieu, P. G., Bryk, A. S., Grunow, A., & Gómez, L. M. (2017). Working to improve: Seven approaches to improvement science in education. *Quality Assurance in Education, 25*(1). Emerald Publishing.

LeMahieu, P. G., & Cobb, P. (Eds.). (2025). *Measuring to improve: Practical measurement to support continuous improvement in education.* Harvard Education Press.

LeMahieu, P. G., & Reilly, E. C. (2004). Systems of coherence and resonance: Assessment for education and assessment of education. *Yearbook of the National Society for the Study of Education, 103*(2), 189–202.

LeMahieu, P. G., & Wallace, R. C., Jr. (1986). Up against the wall: Psychometrics meets praxis. *Educational Measurement: Issues and Practice, 5*(1), 12–16.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education, 30*(1), 109–162.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–8.

Smith, T. (2025). Validity and technical quality of practical measures. In P. G. LeMahieu & P. Cobb (Eds.), *Measuring to improve: Practical measures for improving teaching and learning.* Harvard Education Press.

Solberg, L. I., Mosser, G., & McDonald, S. (1997). The three faces of performance measurement: Improvement, accountability, and research. *The Joint Commission Journal on Quality Improvement, 23*(3), 135–147.

Takahashi, S., Peurach, D. J., Russell, J. L., Cohen-Vogel, L., & Penuel, W. (2022). Measurement for improvement. In D. J. Peurach, J. L. Russell, L. Cohen-Vogel, & W. Penuel (Eds.), *Handbook on improvement research in education.* Rowman & Littlefield.

VOLUME I | SECTION 3

Emerging Technologies in Educational Assessment

Harnessing Emerging Technologies: Innovation *To Assess, To Teach, To Learn*

Eric M. Tucker and Eleanor Armour-Thomas

Recent advances in artificial intelligence (AI) and other technologies are transforming educational assessment—creating opportunities and risks (Ercikan, 2025). Product developers increasingly leverage AI to create interactive tasks and deliver real-time feedback at scale (Ercikan, 2025; Foster & Piacentini, 2025), while generative AI supports content development and automated scoring (Burstein, LaFlair, Yancey, von Davier, & Dotan, 2025). These technologies can make assessment more adaptive, accessible, and formative (Pellegrino, 2025). Yet these innovations also raise questions about whether AI will merely boost efficiency or truly catalyze assessment approaches that serve and enhance learning, consistent with Gordon's (2025) vision. Developers increasingly have rich learner data and methods to detect meaningful patterns (Ercikan, 2025). The challenge is to ensure innovation is guided by values and evidence to advance a learner-centered system (Pellegrino, 2014).

Efficacy, Validity, and Fairness Considerations in AI-Driven Assessments

Ercikan (2025) examines how the integration of AI in educational assessment and measurement offers potential advantages while also raising questions about efficacy, validity, and fairness. She shows how recent advances enable new forms of assessment, from capturing complex competencies to adaptive testing and automated content and scoring. Ercikan concludes by offering guiding questions for researchers and practitioners to critically examine AI's impacts on assessment quality, underscoring that innovation must be paired with careful validation.

Responsible AI for Test Equity and Quality: The Duolingo English Test as a Case Study

Burstein, LaFlair, Yancey, von Davier, & Dotan (2025) argue that specifying and adopting Responsible AI practices is crucial to upholding test quality and test equity in AI-driven assessments (AERA, APA, & NCME, 2014). As a case study, the authors present the Duolingo English Test—an AI-powered, high-stakes language proficiency exam—to illustrate how Responsible AI principles can be translated into practice (ITC & ATP, 2025).

It's Time for a Paradigm Shift in Educational Measurement

Cantor and Felsen (2025) argue that prevailing assessment models rest on flawed assumptions inconsistent with the learning sciences. Developmental science reveals brains remain plastic throughout life, context and experience (via epigenetics) shape learning outcomes, human ability is variable yet widely expandable, and talent is widespread if nurtured under the right conditions (Bransford, Brown, & Cocking, 2000). They advocate realigning measurement with the complex, context-dependent nature of human learning development, they argue, education can better support each student's success.

Toward Personalized, Educative, and Human-Centered Assessment

Taken together, these chapters show how integrating emerging technologies into assessment is part of a larger transformation in how we understand and foster learning (Pellegrino, 2014). Assessment can be made more effective, personalized, and educative—if we intentionally design and deploy AI in alignment with fairness, validity, and human development. These chapters entice our field to prepare us to navigate the opportunities and responsibilities that AI brings to education, so that assessment becomes, in Gordon's words, a means for educating, not just evaluating. The authors invite us to envision a future in which emerging technologies are harnessed through principled innovation to advance practices that are human-centered, just, and cultivate human potential (Gordon, 2020).

References

(AERA, APA, NCME) American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). National Academy Press.

Burstein, J., LaFlair, G. T., Yancey, K., von Davier, A. A., Dotan, R. (2025). Responsible AI for test equity and quality: The Duolingo English Test as a case study. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Cantor, P., & Felsen, K. (2025). It's time for a paradigm shift in educational measurement. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Ercikan, K. (2025). Efficacy, validity and fairness considerations in AI-driven assessments. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Foster, N., & Piacentini, M. (2025). Innovating assessment design to better measure and support learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice, 39*(3), 72–78.

Gordon, E. W. (2025). Series introduction: Toward assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

(ITC & ATP) International Test Commission & Association of Test Publishers. (2022). *Guidelines for technology-based assessment.* https://www.intestcom.org/upload/media-library/guidelines-for-technology-based-assessment-v20221108-16684036687NAG8.pdf

Pellegrino, J. W. (2014). Assessment in the service of teaching and learning: Changes in practice enabled by recommended changes in policy. *Teachers College Record, 116*(11).

Pellegrino, J. W. (2025). Arguments in support of innovating assessments. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Efficacy, Validity and Fairness Considerations in AI-Driven Assessments

Kadriye Ercikan

The rapid advancements and integration of artificial intelligence (AI) are transforming the field of assessment and the science of measurement in unprecedented ways. Particularly in the last two years, generative AI has accelerated significant changes in content development, scoring, interactivity, and personalization (Arslan et al., 2024; Bulut et al., 2024; González-Calatayud et al., 2021; Kyllonen E. et al., 2024; Mao, Chen & Liu, 2024; Zhai & Krajcik, 2024). These AI-driven applications offer opportunities to better align assessments with educational goals by providing opportunities for:

1. Measuring complex competencies essential for success in technologically advanced educational and workplace contexts;
2. Creating engaging, interactive learning and assessment environments that are rewarding for individuals; and
3. Providing targeted feedback to support teaching and learning.

While AI can enhance assessment practices, these new approaches necessitate rigorous evaluation of their efficacy and impact on the validity and fairness of the resulting claims. This perspectives chapter begins by highlighting key opportunities to innovate assessments using AI, followed by illustrative examples. The final section focuses on the critical need for evaluating the efficacy, validity and fairness of AI applications, offering guiding questions for each context.

Opportunities for Innovation in Assessment Using AI

AI offers a range of possibilities to improve assessment quality and efficiency, particularly in measuring complex constructs that have been traditionally difficult to assess through paper-based or linear digital formats (Bennett, 2024; Kyllonen et al., 2024). Notable innovations include:

- **Personalization, interactivity, and adaptivity** to engage test takers more effectively, and optimize performance on assessment;
- **Use of process data** for assessing cognitive processes; and
- **Automation of content creation, scoring, and feedback** to enhance scalability and cost-efficiency.

Below are three examples that illustrate how AI is being used to support learning, personalize assessment experiences, and increase operational efficiency.

Assessment to Support Learning

There is growing interest in using assessments to directly support learning. In a language learning context, AI enables learners to develop their language skills through simulations of workplace tasks. These tasks assess both language proficiency and workplace competencies and are personalized in real time based on learner choices and performance. AI provides immediate feedback and recommendations for further practice (ETS, 2024). Such AI-driven embedded assessment in authentic, real-life learning contexts offers potential for broader educational applications.

Personalizing Assessments

Personalization is a significant AI-enabled opportunity, with the potential to advance fairness and to meet the needs of (neuro)diverse students at scale. By tailoring assessments to align with individuals' linguistic, cultural, and educational contexts—and giving them agency in task selection—personalization provides opportunities to optimize engagement and performance (Arslan et al., 2024; Bennett, 2024). A compelling example is "Context AI," developed by Burcu Arslan and colleagues (Arslan et al., 2024). This tool uses GPT-4 to customize assessment contexts based on student's interests. For example, for a Context AI math item, the student would be allowed to pick an interest area from choices that include things

like football, popular music, or gaming with Roblox. A student selecting Roblox sees an item embedded in the context of this game whereas a student picking other interest areas would have items related to those areas. Prior research shows that the context of test items can significantly influence performance (Ercikan & Solano-Flores, in press) and thus giving students the option to choose the context may be a promising direction.

Increasing Efficiency: Human-AI Collaboration in Scoring

Earliest applications of AI in assessment has been in creating efficiencies in operational large-scale assessments (Williamson et al., 2012) and the most widely used applications of AI currently is in scoring. AI can be used to increase the efficiency of work done by humans in scoring is in two different ways. One is by replacing some of the work done by humans, and the other is to provide tools that help humans to be more efficient and improve the quality of their work. The human-AI collaboration can enhance efficiency through three complementary ways:

1. Verification
2. Contributory scoring
3. Divide and conquer

Verification: AI scores can serve a confirmatory role for human scores by confirming the human scores, and flagging disagreements. They can support quality of human scores by flagging unscorable responses. The flag and review role can also be used to verify AI scores by humans, especially in verifying scoring of "unscorable" responses.

Contributory Scoring: The scores from human raters and AI can be combined in contributory ways where human and AI scores both contribute to the final score. This approach involves both human and AI scores of the same section of the assessment contributing to the reported score. Typically, human and AI both provide holistic scores, with the automated scoring system serving the same role as a second human rater.

Divide and conquer: Humans and AI produce different kinds of information. AI might be used to evaluate specific concrete features of responses, or measure more fine-grained phenomena while humans evaluate broader and more abstract features of meaning and effectiveness. For example, in assessing and providing feedback on writing, AI could be used to provide quantitative feedback, while humans provide qualitative feedback. AI generated feedback can include a summary of the writing, evaluation of cohesion and coherence, language use such as vocabulary as well as mechanics such as spelling. Qualitative feedback might take the form of comments or coaching. In classroom contexts, teachers then can use the AI generated evaluation to personalize and provide feedback to the student. The idea is that teachers can filter the AI feedback and provide nuance around content, both saving time and allowing the teacher to focus on the things that humans do well–interpreting meaning. This kind of efficiency can be useful for teachers as well as for students who can receive the feedback in a more speedy way, than if the teacher had to read and evaluate each essay before they can provide feedback to students about their writing.

Efficacy, Validity and Fairness Considerations in AI-Driven Assessment

The transformative potential of AI in assessment demands thorough evaluations of **efficacy, validity** and **fairness** of claims made based on these assessments. Efficacy refers to whether the application of AI meets the intended goals. Validity is defined as the degree to which claims made based on assessments can be supported by evidence and rationales (Kane, 2006). Fairness refers to the degree to which goals are met across groups and score meaning is consistent for groups (Kane, 2017). Further specifics of these questions are highlighted below for each of the key applications.

Evaluating AI Feedback and Learning Support

When AI is used to provide automated feedback and support learning there are many factors that can influence the impact of the feedback including the nature of the feedback as well as its alignment with the individual learner. Addressing the following questions is central to evaluating efficacy, validity and fairness of this AI application:

- Does the feedback enhance engagement and learning?
- What evidence supports claims of improved engagement (e.g., process or self-report data)?
- Is there evidence of learning based on other assessments?
- Is the evidence supporting improvement in engagement and learning similarly strong for individuals from different backgrounds and contexts?

Evaluating Personalization

Personalization in assessment needs to be evaluated with respect to the degree to which personalization meets its primary goals of enhancing engagement and optimizing performance (Arslan et al., 2024; Bennett, 2024). In addition, when individuals take different forms and formats of assessments in personalized assessment contexts, supporting validity and fairness of assessment claims from these assessments become a key challenge (Sinharay et al., 2025). Empirical evidence is needed to determine whether personalization:

- Increases engagement, performance and measurement precision;
- Improves alignment with learners' interests and backgrounds;
- Supports claims about personalization that can be supported by empirical evidence and rationales; and
- Enhances engagement, performance, and measurement precision consistently for individuals from different backgrounds and contexts.

Evaluating AI-Scoring

Automation for scoring can possibly have the greatest impact on the validity and fairness of claims from assessments. For AI based scoring a thorough evaluation needs to involve how the quality of AI based scores hold up against validity and fairness criteria, as is done for assessments that use human scores. For evaluating fairness of AI based scores we need to evaluate if the validity evidence is consistent across individuals from different contexts.

For **AI-based scoring**, key set of questions for evaluating efficacy, validity and fairness are:

- To what extent are scores consistent with human scoring?
- Is there evidence of construct irrelevant variance?
- Is there evidence of construct under representation?
- Is there evidence of systematic differences between human-AI scoring across groups from different backgrounds and contexts?

Two widely used statistics for evaluating AI-human score agreement are Quadratic Weighted Kappa and Proportion Reduction in Mean Squared Error (PRMSE). High PRMSE values (≥ 0.95) suggest AI scores can be used interchangeably with human ratings. Values ≤ 0.70, however, offer limited validity support for high-stakes use of AI generated scores (McCaffrey et al., 2022).

Concordance between human and machine scores is a requirement, however it is not sufficient for validity of claims based on scores. In addition to human-AI concordance, construct comparability of machine scores and human scores need to be evaluated. Especially given the black-box nature of AI scoring algorithms, we need to evaluate both construct irrelevant variance as well as construct underrepresentation. A special focus on construct underrepresentation is needed to ascertain specific construct components are captured by AI scoring. There are special concerns that the use of AI scoring may contribute to bias. Evaluation of fairness of AI scores need to be considered from the very beginning of the development of AI algorithms (Bennett 2024; Johnson et al., 2025). And checks for fairness must be part of the development of AI scoring models and evaluation. This includes

- Building scoring models using data that represent targeted populations
- Producing evidence that support fair interpretation of scores, as we do with scores from human scoring
- No subgroup difference in distribution of errors from AI scores in comparison with the human scores and
- No differential prediction of human scores (McCaffrey et al., 2022).

Final Note

AI provides many opportunities for assessments to better serve their intended goals such as providing feedback and support for learning, optimization of engagement and performance and increasing efficiency. However, these opportunities present possibilities of significant problems for assessment. Without thorough evaluations, integration of AI in assessment hold possibilities of stereotypical representation of cultural groups, inappropriate and useless feedback, inaccurate scores, and narrowing down of the measurement of the construct being targeted by the assessment. Such risks not only can harm the role of assessment in education but can have important societal impacts. The key questions I presented for each application are intended to highlight **the necessity of empirical research** with a variety of data sources for each application.

References

Arslan, B., Lehman, B., Tenison, C., Sparks, J. R., López, A. A., Gu, L., & Zapata-Rivera, D. (2024). Opportunities and challenges of using generative AI to personalize educational assessment. *Frontiers in Artificial Intelligence, 7,* 1460651.

Bennett, R. E. (2024). Personalizing Assessment: Dream or Nightmare?. *Educational Measurement: Issues and Practice, 43*(4), 119–125.

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., & Morilova, P. (2024). *The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges.* arXiv preprint arXiv:2406.18900.

Ercikan, K., & Solano-Flores, G. (in press). *Socio-cultural context of assessment. In Educational Measurement,* 5th Edition, Linda Cook and Mary Pitoniak, Eds.

ETS. (2024). *Converse Workplace* [mobile app]. Google Play. https://play.google.com/store/apps/details?id=org.ets.convo&utm_source=na_Med

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied sciences, 11*(12), 5467.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education; Praeger.

Kane, M. T. (2012, March 28–29). *Validity, fairness, and testing* [Paper presentation]. Educational Assessment, Accountability, and Equity: Conference on Conversations on Validity Around the World. Teachers College, New York, NY, United States. https://www.tc.columbia.edu/media/media-library-2018/centers-amp-labs/aeri/ conferences-and-forms-/conversations-on-validity-2012-/a283745b-70b8-486c- bdee-4e1f4cda2c48.pdf

Kyllonen, P. C., Sevak, A., Ober, T., Choi, I., Sparks, J., & Fistein, D. (2024). *Charting the Future of Assessments.* ETS Research Report Series, 2024 (1), 1–62.

Mao, J., Chen, B., & Liu, J. C. (2024). Generative artificial intelligence in education and its implications for assessment. *TechTrends, 68*(1), 58–66.

McCaffrey, D. F., Casabianca, J. M., Ricker-Pedley, K. L., Lawless, R. R., & Wendler, C. (2022). Best practices for constructed-response scoring. *ETS Research Report Series*, 2022(1), 1–58.

Sinharay, S., Bennett, R. E., Kane, M., & Sparks, J. R. (2025). Validation for Personalized Assessments: A Threats-to-Validity Approach. J*ournal of Educational Measurement.*

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice, 31*(1), 2–13.

Zhai, X., & Krajcik, J. (Eds.). (2024). *Uses of artificial intelligence in STEM education.* Oxford University Press.

Responsible Artificial Intelligence for Test Equity and Quality: The Duolingo English Test as a Case Study

Jill Burstein, Geoffrey T. LaFlair, Kevin Yancey, Alina A. von Davier, and Ravit Dotan

Abstract

Artificial intelligence (AI) creates opportunities for assessments, such as efficiencies for item generation and scoring of spoken and written responses. At the same time, it poses risks (such as bias in AI-generated item content). Responsible AI (RAI) practices aim to mitigate risks associated with AI. This chapter addresses the critical role of RAI practices in achieving test quality (appropriateness of test score inferences), and test equity (fairness to all test takers)—key principles in this volume. To illustrate, the chapter presents a case study using the Duolingo English Test (DET)—an AI-powered, high-stakes English language assessment. The chapter discusses the DET RAI standards, their development and their relationship to domain-agnostic RAI principles. Further, it provides examples of specific RAI practices, showing how these practices meaningfully address the ethical principles of validity and reliability, fairness, privacy and security, and transparency and accountability standards to ensure test equity and quality.

Introduction

Test quality is achieved through evidence gathering that confirms an assessment's suitability for its intended purpose. *Test equity* is attained when test scores are fair–specifically, they do not favor or disadvantage a particular group. Classical argument-based test validity theory supports test quality and equity by constructing a *chain of inferences* to support test score interpretation. The theory guides the collection and evaluation of evidence to build a validity argument (Chapelle et al., 2008; Kane, 1992). The chain of inferences includes: *domain definition* (task types represent the target domain as defined), *evaluation* (test scores reflect language ability), *generalization* (test scores are reliable), *explanation* (test scores are attributable to the construct), *extrapolation* (test score are related to other language criteria), *utilization* (test scores are interpretable and meaningful for their purpose). While still highly relevant, classical validity theory predates assessments powered by artificial intelligence (AI). To evaluate the validity of interpretations and uses of such assessments, it is essential to evaluate AI capabilities, as they may affect evidence collection, measurement, and ultimately, test quality and equity.

AI-powered assessments are becoming increasingly common, offering many advantages, such as automated scoring of writing and speaking, and efficient creation of larger item banks through automated item generation. However, there are risks. For example, bias in AI-generated item content can impact test-taker outcomes (e.g., Belzak et al., 2025; Johnson et al, 2022); this can, potentially, lead to test inequity and diminished test quality. Therefore, AI-powered assessment calls for alignment with human-centered AI values which are enacted through responsible AI (RAI) guidelines and standards (von Davier & Burstein, 2024; Burstein, 2023; Auernhammer, 2020; The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2017). Though some risks to validity are similar across traditional and AI-based assessments, AI introduces unique risks. Current assessment standards[1] address AI to some extent (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014) (henceforth, AERA&APA&NCME, 2014). However, the expanded use of AI for assessment requires more comprehensive RAI standards and practices to mitigate risks to

1 Standards that were available at the time this chapter was written.

the test validity argument that affect test score interpretation (Association of Test Publishers (ATP), 2024).

This chapter illustrates how RAI assessment standards and practices help to achieve test equity and test quality of AI-powered assessments. To do so, the chapter presents a case study using the RAI standards for the Duolingo English Test (DET)—a digital-first, high-stakes English language assessment. The chapter (i) presents the standards; (ii) explains their development; (iii) validates the DET RAI standards with ethical principles for an industry-agnostic RAI governance framework—the National Institute for Standards and Technology's (NIST) Artificial Intelligence Risk Management Framework (NIST AI RMF) (NIST, 2023); (iv) illustrates their application on the DET, demonstrating how RAI standards and practices support test equity and quality; and, (v) discusses the RAI standards implications and current known limitations for assessment

Background and Related Work

AI[2] has been used for high-stakes assessments for quite a while for automated writing evaluation (AWE). Its use began on first-generation, computer-based assessments[3] (Lottridge et al., 2021; Shermis & Burstein, 2013, 2003; Foltz et al., 1999; Burstein et al, 1998), and speaking evaluation (ASE) (see Zechner & Evanini, 2019). Emerging more recently, *digital-first assessments* (DfA) were "born digital": DfA's were designed to be administered online and leverage AI for test design (e.g., automated item generation), measurement (e.g., automated essay scoring), and security (e.g., plagiarism detection) (see Belzak et al., 2025; Naismith et al., 2025). DfAs are made possible largely due to the availability of generative AI (OpenAI, 2023; Radford et al., 2019), which enables automated item generation at scale (Attali et al., 2022; Khan et al., 2021).

Classical assessment validity (Chapelle et al., 2008; Kane, 2013; 1992) and fairness (Kunnan, 2000) frameworks embody the ethical principles of validity and reliability, and fairness. While developed for paper-and-pencil and first-generation, computer-based assessments, these frameworks have laid the groundwork for

2 Note that not all AI used for assessment is generative AI. In this chapter, we use the term AI to refer to AI and AI-adjacent approaches, including natural language processing (NLP) and statistical modeling approaches.

3 First-generation, computer-based assessments were typically first designed for paper-and-pencil formats and later moved to a computer-delivered format.

modern assessment. However, they do not explicitly address the use of technology for assessment. The AERA, APA, & NCME (2014) *Standards* address automated scoring of written and spoken constructed responses, covering the scope of AI use at the time they were published.

Earlier frameworks and assessment standards also do not address aspects of AI use on tests that may impact test quality, such as the need for AI literacy training of human test developers, or broader societal issues, such as carbon emissions associated with large language model use (Faiz et al, 2024). It is not even clear how these new issues associated with AI use on assessments would be evaluated in the classical validity argument chain of inferences.

Given the growth of AI use on modern digital assessments, guidelines for RAI in learning and assessment have proliferated (such as ATP, 2024; Organization for Economic Cooperation and Development (OECD), 2023; International Test Commission (ITC) & Association of Test Publishers (ATP) (ITC-ATP), 2025; The International Privacy Subcommittee of the ATP Security Committee, 2021; U.S. Department of Education, Office of Educational Technology, 2023). *Guidelines* make high-level recommendations for mitigating AI risks. In contrast, standards translate theoretical principles into practical, actionable guidance (e.g., AERA, APA, & NCME, 2014), offering concrete steps for implementing their underlying ideals. Failing to implement the ideals creates ethical debt that may compromise test quality and equity. This can lead to both short- and long-term harms—such as use of AI-generated content containing hallucinations (e.g., Ji et al., 2023).

The development of RAI assessment standards needs to draw from the extensive set of assessment guidelines and standards. Additionally, it should leverage the rich body of AI ethics literature, which outlines domain-agnostic, ethical principles such as fairness, transparency, explainability, privacy, security, trust, responsibility, justice, and autonomy (Memarian & Doleck, 2023; Floridi & Cowls, 2022; Fjeld et al., 2020; Jobin et al., 2019). Domain-agnostic ethical principles promote human responsibility. When combined with industry-specific standards (e.g., AERA, APA, & NCME, 2014) and guidelines (e.g., OECD, 2023; U.S. Department of Education, 2023; ITC-ATP, 2025), they help ensure alignment with the unique needs of educational assessment, supporting both test equity and quality.

Case Study: The Duolingo English Test RAI Standards

This section presents an overview of the Duolingo English Test (DET), and discusses the DET RAI standards, their development process, and their alignment with a domain-agnostic industry framework. Finally, the case study demonstrates how the systematic application of standards supports assessment quality.

The Duolingo English Test

The DET is a digital-first, high-stakes, computer-adaptive measure of English language proficiency, commonly used for admissions to English-medium higher education institutions (Naismith et al, 2025). It assesses a test-taker's ability to use English language skills that are required for speaking, writing, reading and listening, as well as for integrated skills associated with literacy, conversation, comprehension, and production. Integrated skills require multiple proficiencies, e.g., speaking and listening for conversation.

The DET leverages AI extensively, using automated item generation, automated writing and speaking evaluation, and automated plagiarism detection for test design, measurement, and security, respectively. The DET's test-taker experience also benefits from AI affordances. For instance, the DET's free practice test is made possible by automated item generation (i.e., to generate practice test items) and scoring (i.e., providing an instant score estimate).

The DET employs *human-in-the-loop* (HiTL) AI practices to support test quality and equity. The DET's HiTL approach is consistent with current education and assessment policy, aiming to contribute to the test's equity and quality. Recent education and assessment policies discuss HiTL AI as crucial human oversight at critical decision points (ATP, 2024). HiTL AI is also discussed as a real-time, human-system interaction process, whereby humans provide ongoing input to enhance AI performance (Wang, 2019). The DET leverages human judgment and oversight in test design, measurement, and security. For example, humans review automatically generated content during test design, label training data to build and evaluate writing and speaking models for measurement, and serve as proctors to referee AI-generated plagiarism flags for test security. The case study presented later provides illustrations of these current practices.

The Duolingo English Test Responsible AI Standards

The Four Standards

The DET's Responsible AI standards address test equity and quality through include four standards that represent ethical principles aligned with the test's goals. An overview of the standards is discussed below, and more details are provided in Section 3.3.

1. The ***Validity and Reliability*** standard is crucial to ensure that the test is suitable for its intended purpose. The Validity standard evaluates construct relevance and accuracy, and the Reliability standard focuses on consistency;
2. The ***Fairness*** standard promotes democratization and social justice through increased access, accommodations, and inclusion, representative test-taker demographics, and avoiding algorithms known to contain or generate bias;
3. The ***Privacy and Security*** standard ensures (a) compliance with relevant laws and regulations governing the collection and use of test taker data; (b) ensuring test-taker privacy and (c) providing secure test administration; and,
4. The ***Accountability and Transparency*** standard aims to gain trust from stakeholders through proper governance and documentation of AI used on the test.

Standards Development

Five key activities informed the choice of the four ethical principles used to create the DET RAI standards.

First, we conducted a literature review to identify commonly discussed ethical principles in the context of AI (e.g., Memarian & Doleck, 2023; Floridi & Cowls, 2022; Fjeld et al., 2020; Jobin et al., 2019). The review increased our understanding of which principles were applicable to the DET.

Second, to validate alignment between domain-agnostic, AI governance (e.g., NIST, 2023) and assessment-specific principles, we reviewed well-recognized assessment-specific standards (AERA, & APA, & NCME, 2014) and guidelines (including OECD 2023; U.S. Department of Education, Office of Educational Technology, 2023; and ITC-ATP, 2025).

Third, we consulted a cross-disciplinary group of experts from computational psychometrics, language assessment, law, machine learning, and security within Duolingo, and an external RAI expert from computer science[4].

Fourth, after identifying the four ethical principles, the external RAI expert helped to articulate the rationale and overall goal of each standard, and the more detailed subgoals (i.e., practical implementation of each standard).

Finally, the standards were published as a living document and remain open for public comment.

Connections to NIST AI RMF

The DET RAI Standards were validated against the independent, national, industry-agnostic NIST AI RMF (2023) *trustworthiness characteristics*. Validation with an independent and industry-agnostic ethical framework demonstrates how our standards are aligned with prevailing best practices.

The NIST AI RMF's trustworthiness characteristics are similar to the DET RAI standards' ethical principles in that both identify characteristics of trustworthy AI. The NIST AI RMF emphasizes the following trustworthiness characteristics: Valid and Reliable; Safe, Secure and Resilient; Accountable and Transparent; Explainable and Interpretable; and, Privacy-Enhanced, Fair–with Harmful Bias Managed. Based on the standards development process, the DET RAI standards focus on standards which echo four of these characteristics in Table 1.

4 We thank Dr. Pascale Fung for her expert guidance in the development of the DET RAI Standards (Burstein, 2025)

Table 1

NIST AI RMF trustworthiness characteristics *& DET RAI Standards*

NIST AI RMF trustworthiness characteristic	Description[5]	DET RAI Standards	Description
Valid and Reliable	Ensure objective evidence, fulfilling requirements for intended use. Perform consistently in expected conditions over a period of time.	Validity and Reliability	Ensure that the test is suitable for its intended purpose. Validity standards involve evaluating construct relevance and accuracy, while Reliability standards maintain consistent performance over time.
Fairness with harmful bias managed	Address concerns for equality and equity by addressing issues such as harmful bias and discrimination		Promote democratization and social justice through increased access, accommodations, and inclusion represent test-taker demographics, and avoid algorithms known to contain or generate bias
Privacy Enhanced; Secure and Resilient	Adheres to privacy values such as anonymity, confidentiality Maintain confidentiality, integrity, and availability through protection mechanisms, preventing unauthorized access	Privacy and Security	Ensure (a) compliance with relevant laws and regulations governing the collection and use of test taker data; (b) assurance of test taker privacy and (c) assurance of secure test administration.
Accountable & Transparent	Documents information about an AI system and its outputs for individuals interacting with the system	Accountability and Transparency	Provide thorough documentation and explanations

5 See NIST (2023) for complete descriptions.

How the DET RAI Standards Impact Test Quality and Equity

In this section, we illustrate the application of each of the four DET RAI Standards through practices aligned with their goals and subgoals. The examples show how these practices uphold the standards' ideals and support test quality and equity. To do so, we focus on two DET tasks—the *Interactive Reading* and *Writing Sample* tasks. We briefly describe these task types in Section 3.3.1 (also see Naismith et al., 2025 for more details), and discuss how each of the four standards is applied in test development, measurement, and security (Section 3.3.2).

Task Descriptions: Interactive Reading and Writing Sample

The *Interactive Reading* task is a measure of a test-taker's ability to read in academic contexts. The task contains five different item types, targeting different reading sub-constructs. The item types are: Vocabulary in Context; Text Completion; Reading Comprehension; Main Idea; and Possible Title. The item response formats include multiple choice reading comprehension questions, and a question in which test takers highlight a segment of the text to respond. See Figures 2–6 in the Appendix for screenshots.

The *Writing Sample* is an independent, spontaneous writing task. Test takers receive a prompt, have 30 seconds to prepare, and then have five minutes to write their response. See Figure 7 in the Appendix for a screenshot.

Applying RAI Standards

Examples in this section illustrate how each of the four RAI standards' practices contribute to test quality and equity, referring to the alignment with relevant validity argument interferences.

We begin with the *Validity and Reliability,* and *Fairness* standards 1 and 2. In this context discuss a **six-step RAI process** for DET *task design*. The six-step process also addresses aspects of measurement (e.g., scoring), and security (e.g., item exposure) issues. Steps 1–6 are referred to throughout this section. Discussions for the *Privacy and Security, and Accountability and Transparency* standards 3 and 4 follow.

Validity and Reliability

The Validity and Reliability Standard focuses on test validity (i.e., suitability for its intended purpose) and reliability (i.e., yields consistent results), and has two main goals. The first goal aims to "specify processes required to build a validity argument", and the second goal aims to "evaluate AI used in test item creation, item calibration, and scoring." We illustrate with **four subgoals** from this standard, demonstrating how they contribute to test quality and equity.

Develop a Description for the Test Target Domain–i.e., English Language Proficiency–to Ensure that Test Items Are Aligned with the Domain Being Measured. (Subgoal 1.1.16)

Rationale. This subgoal aligns with the *domain definition* inference and involves defining the construct. Explicitly defining the construct is necessary for the design of any assessment. With regard to RAI, it is essential for construct fidelity when using AI to automatically generate high-quality text passages.

Implementation. Steps 1 and 2 describe the DET's task design process[7], using the Interactive Reading task to illustrate.

Step 1. ***Articulates the target construct, using human subject-matter (assessment science) experts (SME).***

For the Interactive Reading task, the target construct is academic reading. The construct is defined as including a range of reading purposes and cognitive skills categorized into two main areas (Park et al., 2022) important reading skills in university study (Grabe, 2008): Reading for Orientation and *Reading for Information and Argument* (Council of Europe, 2020). Reading for Orientation entails searching for specific information within a text and quickly understanding its general idea with limited information (Giulia Cataldo & Oakhill, 2000; Guthrie, 1988; Guthrie & Kirsch, 1987). Reading for Information and Argument involves understanding main ideas, learning how ideas within a text connect to each other and to the reader's prior knowledge, integrating information from multiple texts or different parts of a long

6 The subgoal nomenclature should be read as follows. The *first numeral* refers to the Standard number, the *second numeral* to the Standard's **Goal number,** and the *third numeral* to the Standard's Subgoal number. For example, **Subgoal 1.1.1** refers to the Validity & Reliability Standard **1,** its Goal **1,** and its Subgoal **1**.

7 The remaining four steps are discussed later under subgoals 1.2.1 and 2.1.3.

text, and using the carefully curated information to interpret the text or perform other tasks (Grabe & Stoller, 2020; Head & Eisenberg, 2009; Thompson et al., 2013).

Step 2. ***Specifies a task and scoring system. This includes AI feature development to operationalize and evaluate the target construct articulated by human SMEs.***

Following the construct definition step, we outline task specifications for passage generation for the Interactive Reading task. These specifications and corresponding scoring systems—including feature development and evaluation—serve to operationalize elements of the target construct, as defined by subject matter experts (SMEs), and support automated passage generation.

The passages are automatically generated to support the Interactive Reading task and fall into two primary text categories: expository and narrative. These categories reflect the target language use domain. Expository texts, such as those found in textbooks (Thompson et al., 2013; Weir et al., 2009) and news articles (Head & Eisenberg, 2009), are particularly relevant for university students. Narrative texts, often used in academic contexts such as ethnographic reports and biographies (de Chazal, 2014), also represent important sources of academic reading.

The automatically generated passages are then used to assess specific aspects of the target construct. *Reading for orientation* is operationalized whereby passages require test takers to demonstrate their comprehension of specific ideas (through text highlighting) and vocabulary knowledge in context (through cloze items). *Reading for information and argument* is operationalized using tasks that require text completion, main idea selection, and passage title identification items.

Evaluate AI Scoring System Accuracy and Fairness, Leveraging Human Expertise (Subgoal 1.1.2)

Rationale. Aligned with Step 2, it is important to evaluate AI scorers during task development to ensure that expected scoring criteria can be satisfied using computationally-derived features. This is especially applicable to production items, such as open-ended writing and speaking items for which the AI-driven features need to align with human scoring rubrics (e.g., grammar error detection, text coherence) (described below). This is the AI analog to human scoring processes, and maps to the explanation inference.

Implementation. AI-driven feature development is most relevant for production tasks. Therefore, we illustrate the implementation of this subgoal using the *Writing Sample* task, where AWE is used to automatically score test-taker responses.

Human experts develop rubrics with criteria for rating writing that is consistent with the Common European Framework of Reference (CEFR) levels and descriptors (Council of Europe, 2020). The rubrics define criteria based on the quality of four writing sub-constructs: *content* (relevance and task achievement), *discourse coherence* (organization and cohesion), *lexis* (including sophistication and correct use of vocabulary in context), and *grammar* (complexity and accuracy). These rubrics are used to train human raters who annotate sizable samples of *Writing Sample* responses. To do this raters use the rubric criteria to assign scores of 1–6, which are aligned with the six CEFR levels (A1, A2, B1, B2, C1 and C2). Consistent with AERA, APA, and NCME (2014) standards and the *evaluation* inference, agreement rates between human raters are monitored to ensure that ratings are reliable and accurate (measurement), as these will be used to build the automated scoring models. Once the AI scoring models are built using a portion of the full human-rated data sets, the remaining portion is used to compute system-human agreement. This is one of the key quantitative evaluation metrics used for AI scoring model development.

Develop (a) explainable scoring methods, and (b) interpretable AI features used for scoring that have clear alignment with domain constructs (Subgoal 1.1.3)

Rationale. Consistent with the AERA&APA&NCME (2014) standards, this subgoal ensures that AI model scores are valid by requiring that scores are explainable, aligning with the explanation inference. The model features discussed below measure various aspects of the domain construct and support score explanation.

Implementation. The *Writing Sample* task is used to illustrate. The DET uses AI models to score open-ended speaking and writing responses. Drawing from the literature on NLP, linguistics, and AWE, human experts identify features to be included in the AI model. These example features represent the Writing Sample task's four sub-constructs.

- **Content:** Inverse document frequency (IDF) weighted word-similarity is used to measure the relevance of the response to the writing prompt (Burstein et al., 1998; Rei & Cummins, 2016).

- **Discourse coherence:** Sentence overlap, coreference counts, and latent semantic analysis (LSA)-based sentence similarity features (Foltz et al, 1998) similar to those implemented in the widely used Coh-Metrix (McNamara & Graesser, 2012) are used to evaluate lexical cohesion—a measure of coherence. In addition, a fine-tuned LLM trained to predict human ratings of coherence is used as a holistic coherence feature (Naismith et al., 2023).
- **Lexis:** Proportion of words by CEFR level (Xia et al., 2016) and differential word use (DWU) (Attali, 2011) are used as measures of lexical sophistication. DWU uses outputs from n-gram classification models to differentiate low and high proficiency test-takers.
- **Grammar:** Tree depth statistics (Schwarm & Ostendorf, 2005) are used as measures of grammatical complexity. Error rates of various grammatical error types, determined through grammatical error correction and classification (Bryant et al., 2017), are used as measures of grammatical accuracy.

To make scoring models explainable, the DET uses SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), whereby response scores are reduced to a sum of feature contributions, which can be aggregated into sub-construct contributions. Using SHAP supports powerful, explainable scoring models with complex, non-linear relationships (such as XGBoost models; See Chen et al., (2016).

Identify AI methods for item creation, leveraging human expertise to efficiently create valid and reliable test items (Subgoal 1.2.1).

Rationale. Large item banks mitigate the long-standing assessment of security issues associated with item exposure and pre-knowledge (Chen et al., 2003; LaFlair et al., 2022; Way, 1998). This subgoal manages automated item generation to efficiently create large item banks using large language models (LLM), such as GPT-4. It aligns with the *domain definition* inference in that it attends to construct relevance.

Implementation. For illustration, we switch back to the Interactive Reading task type, since it has a more complex item generation process than the Writing Sample item type. Steps 3–4 below discuss the implementation.

Step 3. ***Creates prompts that elicit content and questions from a large language model (LLM) (such as, GPT-4). Prompts are expected to meet the specifications for DET tasks. To generate content and texts at scale for the Interactive Reading task, machine learning and assessment scientists collaborate to develop prompts that align with the task specifications*** (See Subgoal 1.1.1).

Step 4. ***Uses LLMs (such as GPT-4) for large-scale generation of questions and content.***

The text types for Interactive Reading (expository and narrative) can be generated at scale via the prompting process (using prompts created in Step 3). One approach to this is to employ in-context learning (Dong et al., 2022), where exemplar texts are provided as part of the prompt submitted to the LLM. In this approach, prototypes of narrative and expository readings are shown to the LLM to generate reading passage outputs. The outputs could be conditioned on any relevant target characteristic, such as a list of STEM subjects or topics.

For *main idea* and *possible title* item types (mentioned earlier), potential **answers** are generated and evaluated based on their similarity to the passage. For *comprehension questions*, the model generates questions and answers, filtering out any content with undesirable characteristics (such as, extreme lengths or poor alignment with the passage). For the *text completion* item type, candidate target sentences are identified based on the probabilistic likelihood of their occurrence in the text. For vocabulary in context items, a different process is used: words for deletion are selected based on likelihood, rank order, syntactic and semantic information, and distance from other elided words. To generate **distractors** for the main idea, possible title, and text completion items, alternative texts, questions, and keys are generated. The keys for the alternate questions are used as distractors for the passage and questions that will be on test. Candidate distractors are selected for human review based on metrics such as vector embedding similarity and LLM log probability. Candidate distractors for vocabulary in context items are then selected from the model's likelihood ranking (targeting lower likelihood values) for the candidate words (Attali et al., 2022). With regard to quality, human review of generated content and items is discussed later as part of Step 5.

Fairness

The Fairness Standard addresses test equity explicitly. It aims to ensure that test takers have equal opportunity to succeed and that AI is free of algorithmic bias. It consists of two main goals. The first goal aims to "specify how the use of AI facilitates test-taker access, accessibility, and inclusion," and the second goal aims to "specify test-taker demographic representation, and algorithms known to contain or generate bias." We focus on two subgoals that address fairness and bias (FAB) review of items, and the mitigation of algorithmic bias.

Develop and apply fairness and bias item review principles for inclusion that eliminate construct-irrelevant barriers and ensure that cultural and linguistic factors do not impede accessibility and inclusion (Subgoal 2.1.5)

Rationale. By focusing on access, accessibility, and inclusion, this standard aims to create a more equitable testing environment for individuals from diverse backgrounds and with varying needs. One of the ways this is achieved is developing and applying item reviews to increase inclusion and eliminate potential biases in automatically-generated test content. This aligns with the explanation inference in that the review manages task (passage) characteristics.

Implementation. Humans review the content and tasks to identify sensitive content and low quality items. The human review process improves the item design process and the prompt development based on human review and feedback. These processes are achieved through Steps 5–6.

Step 5. ***Requires human review of content and tasks.***

To mitigate any negative impact of automatically-generated content that is poorly constructed or distracting, a human review process is conducted to remove such content. This includes both an Item Quality Review (IQR) and a Fairness and Bias (FAB) review. Such reviews are a long-standing tradition in test development (AERA, & APA, & NCME, 2014).

IQR evaluates the content and questions to ensure that they represent the relevant text types (narrative and expository) and sub-constructs targeted by the questions. Factual accuracy is also checked. For both types of reviews, reviewers are trained using in-house materials developed in-house by assessment developers. IQRs are tailored to each task type. Where relevant, they are informed by state-of-the-art

item writing guidelines (Haladyna et al., 2002). F*AB review* ensures that items do not contain content that may introduce culturally sensitive or inaccessible topics that might upset or distract the test-taker and introduce construct-irrelevant variance. What is different about human review of AI-generated from human-generated content is required awareness about potential issues specifically associated with LLM outputs which are unlikely to occur with human-created items, such as LLM hallucinations.

The F*AB guidelines* are an expanded version of Zieky (2015); they are tailored by the DET test developers through regular discussions to avoid inclusion of sensitive content. Additionally, DET test developers survey test takers to understand what types of content they would like to read when taking the test, allowing for test-taker input during test design. While these test-taker surveys do not create fully democratic involvement in the test design process (Jin, 2023; Shohamy, 2001), they incorporate for test taker input about the test content.

Step 6. ***Aims to improve the item design process and the prompt development, using human review and feedback.***

Here, we close the feedback loop between content generation and feedback that is collected from reviewers, weekly research discussions, and test-taker content surveys responses. Information gathered from these sources is used to improve item generation procedures. The information collected through the reviews can be used to improve the prompts for the LLMs, our automated filters of generated content, and our FAB and IQR reviewing guidelines.

This six-step process: 1) aligns the AI-assisted item development process with traditional approaches (e.g., construct articulation, development of specifications, content review) and 2) introduces human evaluation of AI outputs (e.g., content generation and automated scoring) which are likely to have characteristics unique to LLMs.

Evaluate and document demographic representation in data sets used to build AI. Documentation should describe how representative (inclusive) the data are with regard to DET test takers (Subgoal 2.2.1)

Rationale. It is crucial to document and evaluate demographic representation in datasets used for AI-powered assessments. This helps to create inclusive data sets that represent the test-taker population, and mitigate biases in test-taker outcomes, downstream. This is aligned with the evaluation inference as it builds in the awareness of demographic groups.

Implementation. An important part of developing DET's automated writing scorer for the Writing Sample task type is curating human-rated datasets used to train and evaluate scoring models. When building these datasets, Writing Sample responses are selected to include a roughly equal number of test takers identifying as male or female from the seven most common first-language (L1) backgrounds in the test-taker population. L1[8] backgrounds—Arabic, Mandarin Chinese, Telugu, English, Spanish, Gujarati, and Bengali—represent a broad range of language families. This ensures that the model is trained and evaluated on a diverse range of L1 backgrounds, promoting measurement quality.

Evaluate and document bias associated with automatically-generated item content (e.g., Fairness and Bias Review Guidelines), and proficiency measurement (Subgoal 2.2.3)

Rationale. It is essential to evaluate and document known algorithmic bias in AI used in assessment processes, such as test security, design, and measurement. This includes managing potential bias associated with automatically-generated item content and proficiency measurement. This is aligned with the evaluation inference.

Implementation. The DET implements this subgoal for proficiency measurement (scoring) by using Differential Rater Functioning (DRF) analysis (Jin & Eckes, 2021; Myford & Wolfe, 2004) on all scorers for open-ended writing and speaking tasks, including the Writing Sample task. Specifically, these scoring models are evaluated on the representative dataset (mentioned in subgoal 2.2.1) to quantify any bias they may have with respect to sensitive background characteristics (e.g., gender or L1) after controlling response quality. We perform this kind of analysis at both the feature and score level to identify potential differential performance test-takers groups. A similar analysis called differential item functioning (DIF) is used to detect

8 L1 refers to the test-takers' self-reported first language.

bias at the item level (Holland & Wainer, 2012) caused by automatic item generation. The DET conducts differential item functioning (DIF) on its item bank (Belzak et al., 2023), flags items with potential bias, and sends them back for FAB review.

Privacy and Security

The Privacy and Security Standard seeks to ensure that the test administration process is secure, fair, and reliable, while protecting test-taker privacy and preventing cheating. This standard consists of three *goals*. The first goal aims to "specify methods to ensure privacy and security associated with data origin, data collection and processing, and data management". The *second* goal aims to "specify how to maintain test-taker privacy, item security, and test-taker security during test administration". The *third* goal aims to "specify fair and reliable test security proctoring protocols, item pool development, and psychometric procedures for test security." In this section, we highlight a subgoal from this third goal.

Define, document, and implement human-in-the-loop AI proctoring protocols that fairly and reliably identify novel and known cheating behaviors (Subgoal 3.3.1)

Rationale. This subgoal focuses on ways to use AI to identify (evaluate) cheating behaviors, and develop protocols. It supports DET proctors'[9] use of AI-enabled tools to make informed, equitable decisions about cheating behaviors observed on high-stakes assessments. For instance, evidence associated with test takers hiring other people to help them test, or using texts written by others), and, most recently, using AI tools, such as LLMs (e.g., ChatGPT) to generate responses (Khalil & Er, 2023). This is aligned with the *evaluation* inference.

Implementation. We illustrate the implementation of this subgoal on *traditional plagiarism*. Traditional plagiarism is a known problem on high-stakes language assessments (Wang et al., 2019). For example, test takers memorize long, generic essay responses that they superficially adapt to respond to a writing prompt on an assessment. Such cases can often be detected using AI models that quantify feature overlap between texts (Foltýnek et al., 2020). However, it is important to distinguish between *deceitful* plagiarism and *benign* text overlap (Chandrasoma et al., 2004; Pecorari & Petrić, 2014).

9 Note that the DET employs asynchronous proctoring (see Duolingo English Test, 2021)

During proctoring, the DET uses AI to compare test-taker writing responses to a database of relevant Internet content and writing responses from historical DET sessions. Matches are flagged and shown to proctors (See Figure 1 in the Appendix). The DET's plagiarism tool displays the sources where matches were found, and highlights the overlapping text. This demonstrates human-in-the-loop AI (as defined in Section 3.1), as AI tools help human proctors with decision-making. (Note that subgoals within the Accountability and Transparency standard address *AI literacy* requirements to ensure that proctors understand how AI tools work.)

Accountability and Transparency

The Accountability and Transparency Standard seeks to build trust with stakeholders. The standard is satisfied through six goals related to the DET's documentation and dissemination about AI use. (See Burstein, 2025 for details about the six goals). We illustrate standards' application with the first (4.1), second (4.2) and fifth (4.5) goals. The first goal is to "assess how AI processes impact stakeholders". The second goal is related to documenting how "AI is used for building the validity argument, test item creation, test item calibration, and scoring". The fifth goal focuses on "disseminating research about use of AI to various stakeholder communities". Arguably, goals in this standard result in documentation of evidence that supports all inferences in a validity argument, since the documentation offers stakeholders explanation about different aspects AI use for test design, measurement and security.

Document external factors that result in a need to modify AI (Subgoal 4.1.3)

Rationale. External factors can affect the impact of AI use on an assessment. For example, changes in the test-taker population may increase bias as captured by differential item functioning (DIF) or differential rater functioning (DRF).

Implementation. The DET's Analytics for Quality Assurance for Assessment system (AQuAA; Liao et al., 2022) addresses this subgoal. The AQuAA system provides weekly reports on metrics that reflect the quality and comparability of the test scores over time, particularly with respect to shifts in test-taker demographics.

Document AI used for building the validity argument, test item creation, test item calibration, and scoring (Goal 4.2)

Rationale. This documentation ensures that internal stakeholders are fully informed about AI use as they perform their tasks and/or make changes to any part of the test. This ensures that stakeholder actions do not compromise the test's validity, reliability, or fairness.

Implementation. To help satisfy this goal and all its subgoals, the DET documents and controls the use of AI through its Exam Change Proposal (ECP) process. For example, when a new scoring model is developed for task types, such as the Writing Sample, the evidence for the scoring model's validity, reliability, and fairness is collected and documented in an ECP document. Similarly, when a new item is developed for operational use, the evidence that supports the launch of the item is documented in an ECP. Before proposed changes are implemented, the document is reviewed and approved by multiple experts, including DET senior assessment and AI researchers.

Disseminate research about use of AI to various stakeholder communities (Goal 4.5)

Rationale. Outcomes from high-stakes assessments can profoundly impact test takers' educational goals. Test developers should clearly communicate with stakeholder communities about how AI is used across the assessment ecosystem for test design, measurement, and security.

Implementation. To reach different stakeholder audiences, DET researchers regularly disseminate research such as through blog posts, white papers, and peer-reviewed articles. For example, the launch of the Interactive Reading task was accompanied by a white paper describing the task and what it measures (Park et al., 2022), and a subsequent technical, peer-reviewed article describing the procedures for automated generation of the task (Attali et al., 2022).

Limitations and Future Work

Organizational RAI guidelines and standards are not a one-time exercise (PwC, 2024). Organizations that build and deploy AI-powered assessments should commit to integrating RAI into the full assessment ecosystem as a test is developed and deployed. Standards development should evolve in tandem with new versions and applications of AI introduced into the test, while also addressing emerging risks that could affect test equity and quality. Known limitations for the DET RAI standards are discussed here.

AI advances. GPT-4o was released (OpenAI, 2024) only slightly ahead of the time this chapter was being written. GPT-4o is a much more powerful LLM than had previously existed. Its multimodal generation capabilities creates opportunities, such as use for innovative item types. Even more advanced models (e.g., GPT 4.5 and GPT 5.0) have been introduced as we finalized the chapter, and model improvements are likely to continue prior to and following the publication of this chapter. At the same time, the advances present additional risks (such as deep fakes which have implications for test security). To maintain test equity and quality as the technology evolves, assessment developers need to consider how this new technology can be responsibly used for assessment.

Fairness. Fairness issues span across all standards. For example, the use of AI to detect traditional plagiarism[10] was described in the Privacy and Security standard. Since it is acknowledged that AI exhibits biases, it is possible that these detectors introduce biases into the plagiarism evaluation. Given the pervasive nature of fairness issues, one approach to consider is making fairness a cross-standard narrative, decreasing the likelihood that fairness issues fall between the cracks.

Additional RAI Standards. The DET RAI includes four standards: Validity and Reliability, Fairness, Privacy and Security, and Accountability and Transparency. These topics were chosen through a process of literature review, consultation with experts, and internal deliberation. However, naturally, some topics were left out. For example, two important issues the DET RAI Standards do not cover are environmental and labor impacts. The DET's environmental impacts include the carbon emissions and water consumption of the generative AI applications involved in the assessment process, as these impacts are notoriously high (Strubell,

10 The DET recently introduced methods to detect plagiarism behaviors associated with the use of LLMs, and a manuscript describing the methods is in preparation.

Ganesh & McCallum, 2023). The DET is working on estimating the environmental impact of the AI models used on the test. Such impact also includes substitution effects associated with test takers taking the DET instead of an alternative English language proficiency test. For example, if each DET test session were instead replaced by a physical, in-person test session at the closest test center, we estimate that it would require approximately tens of millions of additional kilometers in travel each year. Labor impacts could affect Duolingo's employees as a result of the adoption of generative AI. Currently, there have been no negative impacts as the company's employees have been retrained to integrate generative AI assistance.

The DET's approach to these and other important AI ethics issues is incremental, aiming to increase the scope of the DET RAI standards over time.

5. Implications & Conclusion

Intended implications of this chapter were to increase AI responsibility in assessment with attention to how it may impact *test quality* and *equity*. To do this, we provided a case study that showcases RAI standards customized for an English language proficiency assessment; explains the standards' development process; validates the standards against an AI industry standard–i.e., the NIST AI RMF trustworthiness characteristics; illustrates the standards' implementation; and, facilitates an opportunity for critical professional and public engagement.

The DET RAI standards illustrate one example of how RAI standards and practices can be developed and applied for assessment. Through concrete examples of standards application, this chapter demonstrates how RAI standards contribute to test quality and equity, and ensure that test score interpretations are trustworthy and appropriate. The broader assessment community is invited to consider the DET RAI standards if they choose to develop standards for other assessments.

Acknowledgements

We are grateful to the anonymous reviewers for their insightful comments. Many thanks to our Duolingo English Test colleagues: Mancy Liao, for providing content for plagiarism detection; and, Ed Fu, for content related to environmental impact.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational & psychological testing*. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Association of Test Publishers. (2024). *Creating responsible and ethical AI policies for assessment organizations* (July 12, 2024).

Attali, Y. (2011). *Differential word use for content assessment. Journal of Educational Measurement, 48*(1), 1–22. https://doi.org/10.1111/j.1745-3984.2010.00126.x

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.903077

Auernhammer, J. (2020). Human-Centered AI: The role of human-centered design research in the development of AI. In S. Boess, M. Cheung, & R. Cain (Eds.), *Synergy—DRS international conference 2020* (pp. 1315–1333). https://doi.org/10.21606/drs.2020.282

Belzak, W. C., Baig, B., Naismith, R., Hastings, R., Horie, A. K., LaFlair, G., Liao, M., Niu, C., Shih, Y. S. (2025). *Duolingo English Test: Security and score integrity. DRR-25–01.* Duolingo English Test. https://duolingo-papers.s3.us-east-1.amazonaws.com/reports/DET_Security_Report.pdf

Belzak, W. C., Naismith, B., & Burstein, J. (2023, June). Ensuring fairness of human- and AI-generated test items. In *International Conference on Artificial Intelligence in Education* (pp. 701–707). Springer Nature Switzerland.

Burstein, J. (2025). *The Duolingo English Test Responsible AI Standards* (Duolingo Research Report DRR-25–05, Version 3). Duolingo. https://go.duolingo.com/ResponsibleAI

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. In *36th annual meeting of the Association for Computational Linguistics and 17th international conference on Computational Linguistics: Vol. 1* (pp. 206–210). Association for Computational Linguistics. http://dx.doi.org/10.3115/980845.980879

Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2022). *A theoretical assessment ecosystem for a digital-first assessment—The Duolingo English Test* [Research report]. Duolingo English Test. https://duolingo-papers.s3.amazonaws.com/other/det-assessment-ecosystem-mpr.pdf

Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the association for computational linguistics: Vol. 1. Long papers* (pp. 793–805). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P17-1074

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge. https://doi.org/10.4324/9780203854693

Chandrasoma, R., Thompson, C., & Pennycook, A. (2004). Beyond plagiarism: Transgressive and nontransgressive intertextuality [Publisher: Taylor & Francis]. *Journal of Language, Identity, and Education, 3*(3), 171–193.

Chen, S., Ankenmann, R. D., & Spray, J. A. (2003). Exploring the relationship between item exposure rate and item overlap rate in computerized adaptive testing. *Journal of Educational Measurement*, *40*(2), 129–145. https://doi.org/10.1111/j.1745-3984.2003.tb01100.x.

Chen, T., & Guestrin, C. (2016, August). *Xgboost: A scalable tree boosting system.* In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794).

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment–companion volume*. Council of Europe Publishing. www.coe.int/lang-cefr

de Chazal, E. (2014). *English for academic purposes*. Oxford University Press.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., & Sui, Z. (2022). *A survey for in-context learning.* arXiv preprint arXiv:2301.00234.

Duolingo English Test. (2021). *Duolingo English Test: Security, proctoring, and accommodations*.[White Paper]. https://duolingo-papers.s3.amazonaws.com/other/det-security-proctoring-whitepaper-2021-11.pdf

Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F., & Jiang, L. (2023). *LLMCarbon: Modeling the end-to-end Carbon Footprint of Large Language Models.* The Twelfth International Conference on Learning Representations, https://openreview.net/forum?id=aIok3ZD9to

Fiesler, Casey and Garrett, Natalie. (16 Sept 2020). *Ethical Tech Starts with Addressing Ethical Debt, Wired Ideas:* https://www.wired.com/story/opinion-ethical-tech-starts-with-addressing-ethical-debt/

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI* [Research report]. Berkman Klein Center for Internet & Society at Harvard University. https://dx.doi.org/10.2139/ssrn.3518482

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. In S. Carta (Ed.), *Machine learning and the city: Applications in architecture and urban design* (pp. 535–545). John Wiley & Sons Ltd. https://doi.org/10.1002/9781119815075.ch45

Foltýnek, T., Bjelobaba, S., Glendinning, I., Khan, Z. R., Santos, R., Pavletic, P., & Kravjar, J. (2023). ENAI recommendations on the ethical use of Artificial Intelligence in education. *International Journal for Educational Integrity, 19*(1).

Foltýnek, T., Dlabolová, D., Anohina-Naumeca, A., Razı, S., Kravjar, J., Kamzola, L., Guerrero-Dib, J., Çelik, Ö., & Weber-Wulff, D. (2020). *Testing of support tools for plagiarism detection. International Journal of Educational Technology in Higher Education, 17*(1), 46. https://doi.org/10.1186/s41239-020-00192-4

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2). http://imej.wfu.edu/articles/1999/2/04/index.asp

Giulia Cataldo, M., & Oakhill, J. (2000). Why are poor comprehenders inefficient searchers? An investigation into the effects of text representation and spatial memory on the ability to locate information in text. *Journal of Educational Psychology, 92*(4), 791–799. https://doi.org/10.1037/0022-0663.92.4.791

Grabe, W. (2008). *Reading in a second language: Moving from theory to practice.* Cambridge University Press. https://doi.org/10.1017/CBO9781139150484

Grabe, W., & Stoller, F. L. (2020). *Teaching and researching reading* (3rd ed.). Routledge.

Guthrie, J. T. (1988). Locating information in documents: Examination of a cognitive model. *Reading Research Quarterly, 23*(2), 178. https://doi.org/10.2307/747801

Guthrie, J. T., & Kirsch, I. S. (1987). Distinctions between reading comprehension and locating information in text. *Journal of Educational Psychology, 79*(3), 220–227. https://doi.org/10.1037/0022-0663.79.3.220

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Head, A. J., & Eisenberg, M. B. (2009). *Lessons learned: How college students seek information in the digital age* [Progress report]. University of Washington, The Information School. http://www.ssrn.com/abstract=2281478

Holland, P. W., & Wainer, H. (2012). Differential item functioning. Routledge. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (Version 2). IEEE. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

The International Privacy Subcommittee of the ATP Security Committee. (2021, July 6). *Artificial intelligence and the testing industry: A primer.* Association of Test Publishers.

International Test Commission & Association of Test Publishers (2025). *Guidelines for technology-based assessment.* Association of Test Publishers; International Test Commission.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.

Jin, Y. (2023). Test-taker insights for language assessment policies and practices. *Language Testing, 40*(1), 193–203. https://doi.org/10.1177/02655322221117136

Jin, K.-Y., & Eckes, T. (2021). Detecting differential rater functioning in severity and centrality: The dual DRF facets model. *Educational and Psychological Measurement 82*(4), 757–781. https://doi.org/10.1177/00131644211043207

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement*, *59*(3), 338–361.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://psycnet.apa.org/doi/10.1037/0033-2909.112.3.527

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73. https://doi.org/10.1111/jedm.12000

Khalil, M., & Er, E. (2023, June). Will ChatGPT Get You Caught? Rethinking of Plagiarism Detection. In *International Conference on Human-Computer Interaction* (pp. 475–487). Springer Nature Switzerland.

Khan, S., Hamer, J., & Almeida, T. (2021). Generate: A NLG system for educational content creation. In I.-H. Hsiao, S. Sahebi, F. Bouchet, J. -J.Vie (Eds.), *Proceedings of the 14th international conference on educational data mining* (pp. 736–740). Educational Data Mining.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Studies in language testing: Vol. 9. Fairness and validation in language assessment: Selected papers from the 19th Language Testing colloquium, Orlando, Florida* (pp. 1–14). Cambridge University Press.

Liao, M., Attali, Y., Lockwood, J. R., & von Davier, A. A. (2022). Maintaining and monitoring quality of a continuously administered digital assessment. *Frontiers in Education, 7*. https://doi.org/10.3389/feduc.2022.857496

Lottridge, S., Godek, B., Jafari, A., & Patel, M. (2021). *Comparing the robustness of deep learning and classical automated scoring approaches to gaming strategies* [Technical report]. Cambium Assessment Inc.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems: Vol. 30.* Curran Associates, Inc.

McNamara, D. S., & A. C. Graesser. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing." *Applied natural language processing: Identification, investigation and resolution* (pp. 188–205). IGI Global.

Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in Artificial Intelligence (AI), and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, *5*. https://doi.org/10.1016/j.caeai.2023.100152

Myford, C., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189–227.

Naismith, B., Cardwell, R., LaFlair, G., Nydick, S., & Kostromitina, M. (2025). *Duolingo English Test: Technical manual* (Duolingo Research Report). Duolingo. https://go.duolingo.com/dettechnicalmanual

Naismith, B., Mulcaire, P., & Burstein, J. (2023). Automated evaluation of written discourse coherence using GPT-4. *In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (BEA 2023, pp. 394–403). Toronto, Canada: Association for Computational Linguistics.

National Institute of Standards and Technology (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

OpenAI (2024). https://openai.com/index/hello-gpt-4o/

OpenAI (2023). *GPT-4 Technical Report*. https://arxiv.org/pdf/2303.08774.pdf

OECD (2023). *Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI* (No. 349). OECD Publishing. https://doi.org/10.1787/2448f04b-en

Park, Y., LaFlair, G. T., Attali, Y., Runge, A., & Goodwin, S. (2022). *Interactive reading—The Duolingo English Test* [White paper]. Duolingo English Test. https://doi.org/10.46999/RCXB1889

Pecorari, D., & Petrić, B. (2014). Plagiarism in second-language writing. *Language Teaching, 47*(3), 269–302. https://doi.org/10.1017/S0261444814000056

PriceWaterhouseCoopers (PwC) (2024). *PwC's 2024 US Responsible AI Survey.* https://www.pwc.com/us/en/tech-effect/ai-analytics/responsible-ai-survey.html

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 1–18.

Rei, M., & Cummins, R. (2016). Sentence similarity measures for fine-grained estimation of topical relevance in learner essays. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the 11th workshop on innovative use of NLP for building educational applications* (pp. 283–288). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W16-0533

Schwarm, S., & Ostendorf, M. (2005). Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (ACL'05), pages 523–530, Ann Arbor, Michigan. Association for Computational Linguistics. https://doi.org/10.17705/1thci.00131

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation.* Routledge.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective.* Routledge.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language testing, 18*(4), 373–391. https://journals.sagepub.com/doi/abs/10.1177/026553220101800404

Strubell, E., Ganesh, A., & McCallum, A. (2019, July). *Energy and policy considerations for deep learning in NLP.* In Proceedings of the Annual Meeting of the Association for Computational Linguistics, (pp. 3645–3650).

Thompson, C., Morton, J., & Storch, N. (2013). Where from, who, why and how? A study of the use of sources by first year L2 university students. *Journal of English for Academic Purposes, 12*(2), 99–109. https://doi.org/10.1016/j.jeap.2012.11.004

U.S. Department of Education, Office of Educational Technology (2023). *Artificial intelligence and future of teaching and learning: Insights and recommendations* [Report]. https://www2.ed.gov/documents/ai-report/ai-report.pdf

Von Davier, A., & Burstein, J. (2024). *AI in the Assessment Ecosystem: Implications for Fairness, Bias, and Equity.* In Artificial Intelligence in Education: The Intersection of Technology and Pedagogy, Springer Nature Switzerland.

Wang, G. (2019, October 20). H*umans in the Loop: The Design of Interactive AI Systems:* https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems

Wang, X., Evanini, K., Mulholland, M., Qian, Y., & Bruno, J. V. (2019). *Application of an Automatic Plagiarism Detection System in a Large-scale Assessment of English Speaking Proficiency.* Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, 435–443.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17–27. https://doi.org/10.1111/j.1745-3992.1998.tb00632.x

Weir, C., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2009). *The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university* (vol. 9; pp. 97–156). British Council; IELTS Australia.

Xia, M., Kochmar, E., & Briscoe, T. (2016). Text readability assessment for second language learners. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the 11th workshop on innovative use of NLP for building educational applications* (pp. 12–22). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W16-0502

Zechner, K., & Evanini, K. (Eds.). (2019). *Automated speaking assessment: Using language technologies to score spontaneous speech.* Routledge.

Zieky, M. J. (2015). Developing fair tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 81–99). Routledge.

Appendix

Figure 1

Screenshot of the proctor interface.

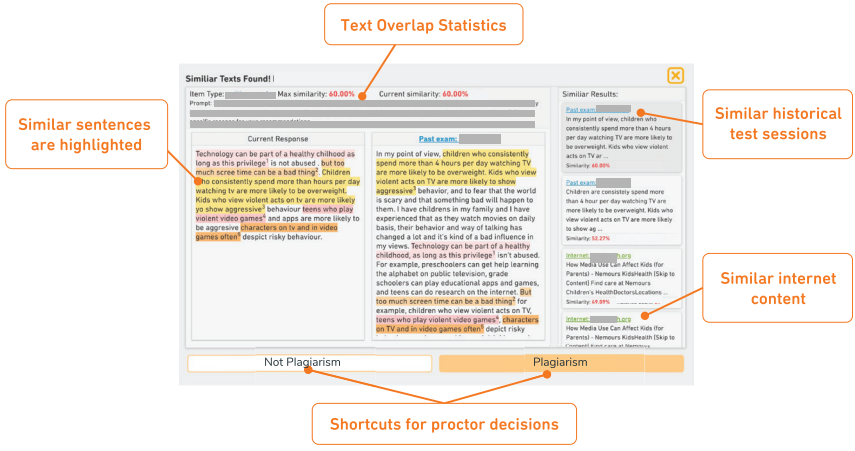


Figure 2

Interactive Reading: Vocabulary in Context (Cloze)

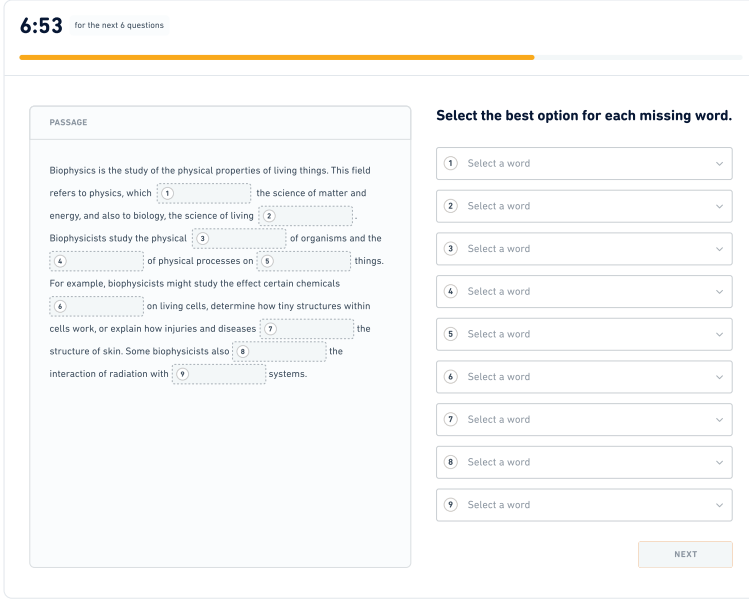


Figure 3
Interactive Reading: Text Completion

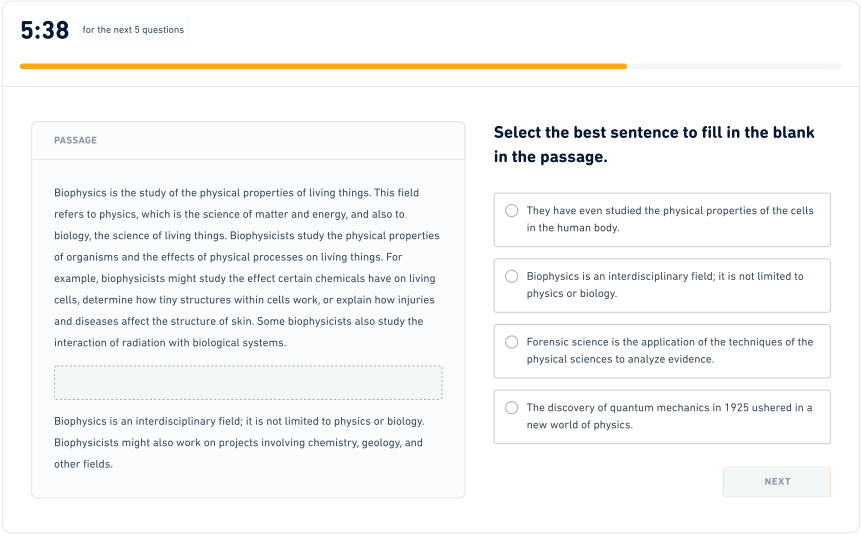


Figure 4
Interactive Reading: Comprehension Questions

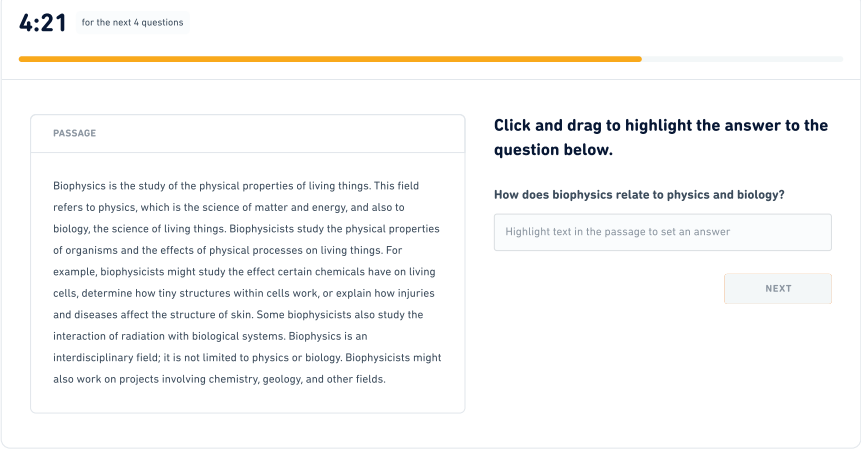


Figure 5
Interactive Reading: Main Idea

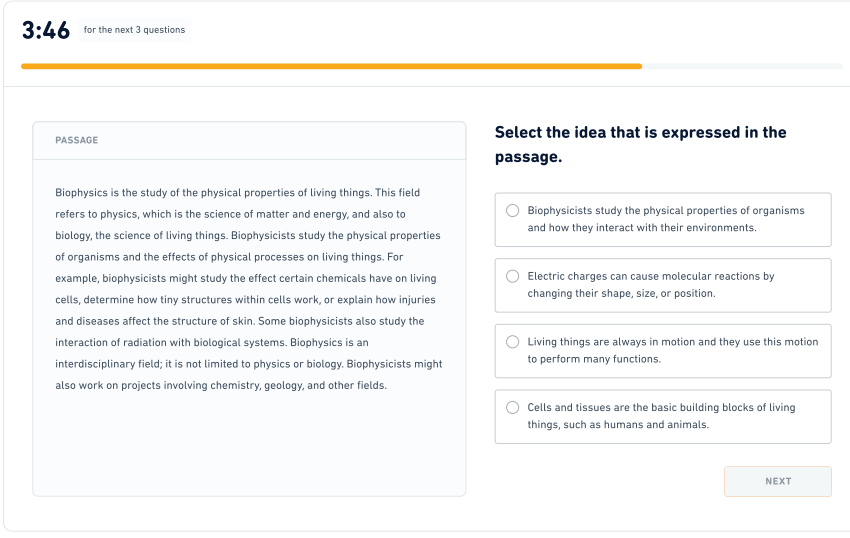


Figure 6
Interactive Reading: Possible Title

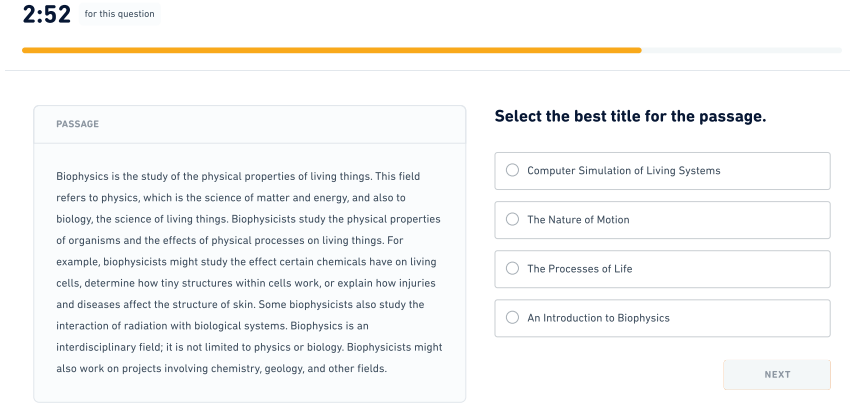
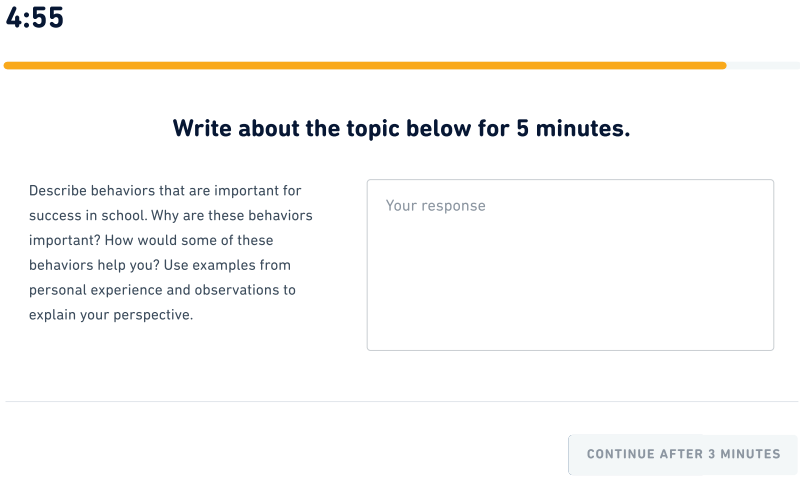


Figure 7
Writing sample



It's Time for a Paradigm Shift in Educational Measurement

Pamela Cantor and Kate Felsen

Science and History

For nearly 1,500 years, from the time of Ptolemy until the 17th century, most people believed the Earth was the center of the universe and the planets and the sun revolved around it. Maps and almanacs depicted this geocentric system that seemed logical at the time. Logical, but also dead wrong. And that meant early navigation tools were wrong, so some ships never reached their destinations.

Then, in 1543, Nicolaus Copernicus proposed that the sun was the center of the universe and built mathematical models accordingly (Copernicus, 1543). Over the next century and a half, astronomers, including Galileo Galilei, refined telescopes and designed better navigation tools. Marine chronometers, sextants and octants improved and so did travel by ship. As the Copernican heliocentric view took hold, understanding of the nature of the universe grew.

In 1928, Alexander Fleming discovered that a mold in one of his petri dishes—something he called penicillin—could kill bacteria (Fleming, 1929). At that time many still believed that miasma or foul air caused disease. Few grasped the concept of germ theory posited by scientists such as Louis Pasteur and Robert Koch, that particles not visible to the human eye, whether bacteria or viruses, were wreaking havoc on human health (Koch, 1882; Pasteur, 1878).

But when penicillin was used to treat soldiers' infected wounds during World War II and saved millions of lives, germ theory took hold. This led to the development of new antibiotics, vaccines and antiseptics. Coupled with better optics in the form of microscopes that could observe pathogens, more accurate diagnoses and better interventions became possible. The field of medicine was changed forever.

The Structure of Scientific Revolutions

In 1962, Thomas Kuhn wrote a landmark essay called "The Structure of Scientific Revolutions" where he described the conditions that can bring about a seismic change, what he called a paradigm shift, where theories and applications are given up in favor of new and improved ones that lead to more and more coherent and innovative models (Kuhn, 1962).

Kuhn stated that three conditions precede scientific revolutions:

1. A flawed theory that leads to what he called mistakes, applications that are incorrect or insufficient and pile up because they are not based in solid science. These mistakes cause mounting frustration and failure, making anything other than incremental change impossible;
2. New instrumentation that enables new mathematical models and algorithms that better account for what is observed; and
3. An alternate theory that better explains what is observed in practice and in human beings.

Based on Kuhn's rubric, we are living in a time that is ripe for a scientific revolution in how we understand human growth and learning, human performance and potential, and how we measure it.

Flawed Theory

Flawed theories are not only alive and well today, they are also foundational principles for learning and measurement systems in the United States. One of the most damaging is the idea that intelligence is fixed, that our brains cease to develop at the time we start kindergarten. That is not true. The human brain is malleable, even highly plastic, and can grow, heal and change throughout our lifetimes (Cantor et al., 2019). All the more reason to measure an individual over time, rather than once or twice a year.

Consider the persistent myth about the power of genes. Many people still believe that genes are the drivers of who we become, including our intelligence. That is false. What matters most is not whether we have a particular gene, but what our genes are doing. Genes can be turned on and off like a dimmer switch. Context–the environments, relationships and experiences in our lives – controls this switch (Cantor & Osher, 2021).

Our genome is not just something we are born with. It develops, as we do, over our lifetimes because of the influence of context. The field of epigenetics describes this phenomenon; "epi" means what sits above—in this case, above the gene—and influences or drives its expression. The genome you have at birth is not the same as the one you will have when you die or the one you will pass on to your children (National Scientific Council on the Developing Child, 2010). David Moore's 2015 book, *The Developing Genome* describes this developmental process in fascinating detail (Moore, 2015).

A third flawed theory is that talent is scarce and distributed like a bell curve, with most of us falling in the middle and a smaller number on each of the tail ends. The notion that only a select few are capable of high levels of achievement has been hard to shake, leading to deeply ingrained inequities in education and employment that constrain people from achieving what they might under more equitable circumstances. This statistical model has been disproven for decades, most powerfully in Stephen Jay Gould's *The Mismeasure of Man* (Gould, 1981). The truth is that human variation is the norm in human development, not the exception (Rose, 2016). No group, race, or ethnicity has more talent or intelligence or more potential for developing it. Talent is everywhere. We don't always look for it, see it, or recognize it, but it is out there, and there are many pathways to develop it.

A fourth flawed theory concerns potential or developmental range. What's false is that human potential is knowable in advance, that we can test for it. What's true is that the fullest expression of a human being's potential depends on contextual factors to reveal and build it (Cantor et al., 2021). Every human skill has a developmental range of performance, and for the most part we all live without knowing what the upper end of that range might be.

By contrast, when a context is designed to reveal capabilities—in learning, in sports, in the workplace—we can discover potential and grow it. Kurt Fischer called this kind of context a "constructive web" in his 2007 paper, "The Dynamic Development of Action and Thought" (Fischer, 2007). When he wrote it, Fischer surely didn't have the 2024 Paris Olympic Games in mind, but every single story of athletic performance we witnessed there is a story of the bi-directional connection between developmental range and the power of context.

Take 9-time Olympic gold medalist Katie Ledecky. In a poolside interview with NBC Sports on July 31, 2024 she described what she was thinking as she pulled away from the field in the 1500 meter freestyle: "I let my mind wander during the race, thinking of all the people that have trained with me, just kind of saying their names in my head and thinking about them" (NBC Sports, 2024). Ledecky trained in contexts that included relationships with teammates, coaches, and physiotherapists who knew how to bring out her best and the best in each other. And she was reliant on data from underwater cameras about each stroke, kick, and training regimen to do it. The full web of experiences constructed around this one human has helped her become one of the greatest swimmers of all time.

New Developmental Insights Lead to a Dynamic Approach to Measurement-3D Measurement

Like Copernicus, who wanted to understand the whole universe, measurement in medicine seeks to understand the whole human being. To do so, medical professionals use three forms of measurement: population, differential, and person-specific (Parrish, 2010). Medicine also assesses contextual factors around each patient, including nutrition, sleep patterns, and living conditions (Cook et al., 2023).

When a doctor takes your blood and compares your white blood cell count to large populations to tell if yours is within the normal range, that is an example of population measurement. In sports, a Nordic skiing or biathlon coach might want to measure an athlete's resting heart rate and compare it to the general population. If the athlete is fit, it should be lower than the population average. If it's not, it might be a sign of inadequate conditioning or maybe illness.

Differential measurement is about comparisons, comparing something about you, such as high blood pressure, to people who are like you, say women in their 50s. When it comes to performance, we need to compare your score to the top scores to know how to shape the goal you are shooting for—such as VO2Max in sports with sprints or a five on an Advanced Placement exam.

The third form of measurement is person-specific, looking at an individual's progress over time. In medicine, this might mean looking at the range of motion in your wrist or ankle four weeks after surgery and comparing it two months later. In music, teachers of brass or wind instruments such as the trumpet or bassoon might measure lung capacity or diaphragm strength in their pupils, then tailor

individual breathing exercises for them. Weeks later, they might repeat the student's measurements to see if the regimen has resulted in positive changes. In athletics, runners training for the 100-meter dash want to know if their training regimen is helping to lower their time from one meet to the next.

In most learning and workforce settings, however, we don't measure in three dimensions. We rely heavily on standardized assessments, many of which are interpreted using norm-referenced frameworks—comparing individuals to a generalized group rather than to relevant peers or performance benchmarks. This approach often ignores individual context, which is crucial for personalizing learning and growth (Cantor et al., 2021; Immordino-Yang et al., 2023).

No surprise, aspiring Olympians are highly attuned to how they feel physically and mentally. It's also true that they understand acutely that the interaction between their bodies and who and what's around them impacts whether they will stand on top of the podium or not. In other words, athletes perform differently at different times and in different contexts, and they need to know why to optimize performance. We do not do this for learning, at least not yet.

Educational measurement should be done in such a way that we learn more about the fit between an individual student and the learning context. This is doable today. It is fit that amplifies purpose and confidence. It is fit that primes performance. It is fit that produces cures. It is fit that unlocks human potential. The Rosetta Stone for measurement is about how close we can get to measuring the individual, understanding and measuring the context, and measuring the fit between the two (Cantor et al., 2021). If this existed today, all learners, and all of us, would understand what we need to reach the top of our developmental range, and what, in our community, school, work, team, troop, would enable us to perform at our best.

New Paradigms

Thomas Kuhn helped us to see that new models and innovations can flow out of changes in theories, mathematics, instrumentation, and methods. The mathematical psychologist Peter Molenaar has moved us to understand that averages do not tell us enough about individuals. In his 2004 paper, "A Manifesto on Psychology as Idiographic Science: Bringing the Person Back into Scientific Psychology, This Time Forever", he demonstrated that it is possible to measure human development in context and even gain insights about the specific impact of a specific aspect of context on a specific person at a specific moment in time (Molenaar, 2004).

Dr. Eli Van Allen is doing something similar at the level of the cell. He leads Clinical Computational Oncology at the Dana Farber Cancer Institute in Boston where his approach to data analytics has led to innovations. One of them is the ability to describe the relationship between a cancer cell and its context (Van Allen et al., 2016). Today, targeted immunotherapies can make changes to the context of individual cancer cells in a patient and destroy those cells without damaging surrounding tissue. This discovery is revolutionizing the treatment of highly lethal cancers.

Emerging technologies and dynamic measurement methods can help us reshape how we understand and nurture talent, learning, and performance. The integration of artificial intelligence (AI) into assessment methodologies is giving us new insights into how individuals learn, develop skills, and optimize their abilities over time. Along with machine learning and computational modeling, AI produces:

- **Personalized learning environments** that adapt to individual strengths, weaknesses, and learning styles
- **Real-time feedback mechanisms** that help learners and educators make data-driven decisions about instructional strategies.
- **Advanced predictive analytics** that identify potential barriers to success and recommend targeted interventions

By integrating AI-driven assessment tools, educators and employers can gain deeper insights into an individual's learning patterns, strengths, and areas for improvement. AI-powered adaptive learning platforms are already demonstrating success in tailoring instructional content to individual needs, optimizing the pace and complexity of material based on a learner's progress. In workforce development, AI-driven skills mapping and competency-based assessments can help match individuals to roles that align with their strengths and career aspirations. And by embracing dynamic measurement systems, we can move away from rigid evaluation models and toward a more personalized and adaptive approach to human development.

Can anyone become a gymnast like Simone Biles, a cellist like Yo-Yo Ma or a poet like Amanda Gorman? Likely no. But none of us can know what is encoded in our DNA without experiences that bring that code to life and without measurement capability that can see potential as it is unfolding. All of us could discover the top of our developmental range if only we had the tools to see it.

References

Cantor, P., Gomperts, N., Lerner, R. M., Pittman, K., & Chase, P. (2021). *Whole-child development, learning, and thriving: A dynamic systems approach.* Cambridge University Press.

Cantor, P., & Osher, D. (Eds.). (2021). *The science of learning and development: Enhancing the lives of all young people.* Routledge.

Cantor, P., Osher, D., Berg, J., Steyer, L., & Rose, T. (2019). Malleability, plasticity, and individuality: How children learn and develop in context. *Applied Developmental Science, 23*(4), 307–337. https://doi.org/10.1080/10888691.2017.1398649

Cook, C. E., Bailliard, A., Bent, J. A., Bialosky, J. E., Carlino, E., Colloca, L., Esteves, J. E., Newell, D., Palese, A., Reed, W. R., Vilardaga, J. P., & Rossettini, G. (2023). An international consensus definition for contextual factors: findings from a nominal group technique. *Frontiers in Psychology, 14,* Article 1537242.

Copernicus, N. (1543). *De revolutionibus orbium coelestium* [On the revolutions of the celestial spheres]. Johannes Petreius.

Fischer, K. W. (2007). Dynamic development of action and thought: The role of the constructive web. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (6th ed., pp. 313–399). Wiley.

Fleming, A. (1929). On the antibacterial action of cultures of Penicillium, with special reference to their use in the isolation of B. *influenzae. British Journal of Experimental Pathology, 10*(3), 226–236.

Gould, S. J. (1981). *The mismeasure of man.* W.W. Norton.

Immordino-Yang, M. H., Nasir, N. S., Cantor, P., & Yoshikawa, H. (2023). Weaving a colorful cloth: Centering education on humans' emergent developmental potentials. *Review of Research in Education, 47*(1), 1–45. https://doi.org/10.3102/0091732X231223516

Koch, R. (1882). Die Ätiologie der Tuberkulose [The etiology of tuberculosis]. *Berliner Klinische Wochenschrift, 19*(15), 221–230.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* University of Chicago Press.

Molenaar, P. C. M. (2004). *A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. Measurement, 2*(4), 201–218.

Moore, D. S. (2015). *The developing genome: An introduction to behavioral epigenetics.* Oxford University Press.

National Scientific Council on the Developing Child. (2010). *Early experiences can alter gene expression and affect long-term development* (Working Paper No. 10). Center on the Developing Child, Harvard University.

NBC Sports (2024, July 31). *Katie Ledecky Olympic 1500m freestyle post-race interview* [Television broadcast]. NBC Sports.

Parrish, R. G. (2010). *Measuring population health outcomes. Preventing Chronic Disease, 7*(4), A71. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2901569/

Pasteur, L. (1878, April 29). *La théorie des germes et ses applications à la médecine et à la chirurgie* [Speech]. French Academy of Sciences, Paris, France.

Rose, T. (2016). *The end of average: How we succeed in a world that values sameness.* HarperOne.

Van Allen, E. M., et al. (2016). Genomic Approaches to Understanding Response and Resistance to Immunotherapy. *Clinical Cancer Research, 22*(23), 5642–5650.

Looking Back, Moving Forward: Reflections on the Foundations for Assessment in the Service of Learning

Eleanor Armour-Thomas, Sheryl L. Gómez, and Eric M. Tucker

Volume I invites us to approach assessment as integral to pedagogy—a continuous dialogue that guides learning—so it functions as an instrument for understanding—and not only as an instrument for determining status. Professor Edmund W. Gordon's charge, in this regard, is clear: rebalance assessment's role and embrace it as an approach for cultivating intellective competence—as a relational, pedagogical enterprise (Gordon, 2025). We must measure what matters—valuing transferable competencies; innovate how we assess—using authentic, interactive tasks grounded in how people learn; and measure in a manner that is useful and usable while upholding technical requirements for validity, fairness, scientific soundness, transparency and credibility.

Designing Assessment in the Service of Learning: Principles and Frameworks

Section 1 argues that rebalancing the focus of assessment begins with principled design. What we intend to infer must govern how evidence is generated, scored, and interpreted (AERA, APA, & NCME, 2014; Pellegrino, Chudowsky, & Glaser, 2001). Grounded in how people learn, principled design aligns claims, observations, and interpretations, embeds tasks in instruction, and meets heightened obligations for transparency, reliability, and fairness (Darling-Hammond & Adamson, 2014). Section 1 offers architecture as well as aspiration: a coherent infrastructure linking cognitive models, measurement, and use, including while designing and developing performance task-, simulation-, game-, or AI-powered assessment.

Grounding Assessment in How People Learn: Scientific Foundations

Section 2 grounds assessment firmly in contemporary science of learning and development: to serve learning, assessment must track understanding as it actually unfolds—dynamic, relational, and situated—and return useful and usable evidence to educators and learners (Nasir, Lee, Pea, & McKinney de Royston, 2020). It adopts an individual-in-context stance, recognizing that cognition grows through interactions with social, cultural, and material environments. The pedagogical throughline is clear: evidence should arise inside instruction, through discipline-specific tasks that make thinking visible and provide timely feedback so teachers can adjust and students can take the next step (Black & Wiliam, 1998). Efficacy is designed from the start by accommodating and recognizing the strengths of human variation and offering multiple, authentic ways to demonstrate competence (Armour-Thomas, McCallister, Boykin, & Gordon, 2019). The section advances practical measurement for improvement—frequent, purpose-built checks that monitor growth over time—and argues for coherence and balance across levels of the education system (LeMahieu& Cobb, 2025; Marion, Pellegrino & Berman, 2024).

Harnessing Emerging Innovations and New Possibilities

Section 3 adopts critical optimism about emerging technologies, arguing that innovation must be purpose-first and human-centered. Artificial intelligence, simulations, and game-based environments can have the affordances to elicit richer evidence of complex skills, personalize challenge and support, and deliver timely feedback, making assessment a seamless part of learning rather than a separate event. Such tools demand strong validity arguments, vigilance for bias, transparent operation, and firm commitments to privacy, data stewardship, and fair access.

The Call to Operationalize Pedagogical Analysis

Through a new paradigm Gordon calls Pedagogical Analysis, he champions the importance of collecting and analyzing pedagogical evidence—complementing status measures—so we can understand and improve assessment–teaching–learning transactions. Status evidence remains vital, but on its own is insufficient to "inform and improve teaching and learning processes and outcomes" (Gordon Commission, 2013, p. vii). that catalyze and cultivate intellective competence in all learning persons.

Conclusion: From Foundations to Frameworks to Transformation

Volume I seeks to understand assessment as principled design and a relational act whose purpose is to advance teaching and learning. This means building systems that do more than register outcomes—they must illuminate how students think, struggle, and progress, so that teaching with supports can be responsive to their assets, needs and interests. Volume II advances the 'how,' bridging design principles and technical innovation; Volume III demonstrates those principles in action across varied contexts. If we measure what matters, innovate how, and measure well, assessment becomes an ally of teaching and a lever for learning—pursuing Gordon's vision of assessment in the service of learning.

References

(AERA, APA, NCME) American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Armour-Thomas, E., McCallister, C., Boykin, A. W., & Gordon, E. W. (Eds.). (2019). *Human variance and assessment for learning.* Third World Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Darling-Hammond, L., & Adamson, F. (2014). *Beyond the bubble test: How performance assessments support 21st-century learning.* Jossey-Bass.

Gordon, E. W. (2025). Series introduction: Toward assessment in the service of learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

LeMahieu, P., & Cobb, P. (2025). Practical Measurement for Improvement: Foundations, Design, Rigor. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst Libraries.

Marion, S. F., Pellegrino, J. W., & Berman, A. I. (Eds.). (2024). *Reimagining balanced assessment systems.* National Academy of Education.

Nasir, N. S., Lee, C. D., Pea, R., & McKinney de Royston, M. (Eds.). (2020). *Handbook of the cultural foundations of learning.* Routledge.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press. https://doi.org/10.17226/10019

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Series Contributors

Sergio Araneda, University of Massachusetts Amherst

Eleanor Armour-Thomas, Queens College, City University of New York (Emeritus)

Aneesha Badrinarayan, Education First

Eva L. Baker, University of California, Los Angeles, Center for Research on Evaluation, Standards, & Student Testing (CRESST), School of Education & Information Studies (SE&IS)

Hee Jin Bang, Age of Learning

Héfer Bembenutty, Queens College, City University of New York

Randy E. Bennett, ETS, Research Institute

Anastasia Betts, Learnology Labs

Mary K. Boudreaux, Southern Connecticut State University

Susan M. Brookhart, Duquesne University

Carol Bonilla Bowman, Ramapo College of New Jersey

Jack Buckley, Roblox

Jill Burstein, Duolingo, Inc.

Pamela Cantor, The Human Potential L.A.B.

Jennifer Charlot, RevX

Gregory K. W. K. Chung, University of California, Los Angeles, Center for Research on Evaluation, Standards, & Student Testing (CRESST), School of Education & Information Studies (SE&IS)

Paul Cobb, Vanderbilt University

Kimberly Cockrell, The Achievement Network, Ltd.

Kelly Corrado, PBS KIDS

Danielle Crabtree, University of Massachusetts Amherst

Linda Darling-Hammond, Learning Policy Institute

Jacqueline Darvin, Queens College, City University of New York

Girlie C. Delacruz, Northeastern University

Clarissa Deverel-Rico, BSCS Science Learning

Kristen Eignor DiCerbo, Khan Academy

Ravit Dotan, TechBetter LLC

Kerrie A. Douglas, Purdue University

Kadriye Ercikan, Educational Testing Service

David S. Escoffery, Educational Testing Service

Series Contributors (continued)

Carla M. Evans, National Center for the Improvement of Educational Assessment

Howard T. Everson, Graduate Center, City University of New York

Cosimo Felline, PBS KIDS

Kate Felsen, The Human Potential L.A.B.

Tianying Feng, University of California, Los Angeles, Center for Research on Evaluation, Standards, & Student Testing (CRESST), School of Education & Information Studies (SE&IS)

Natalie Foster, Organisation for Economic Cooperation and Development (OECD)

James Paul Gee, Arizona State University (Emeritus)

Sheryl L. Gómez, The Study Group

Edmund W. Gordon, Teachers College, Columbia University (Emeritus); Yale University (Emeritus)

Sunil Gunderia, Age of Learning

Laura S. Hamilton, National Center for the Improvement of Educational Assessment

Emily C. Hanno, MDRC

John Hattie, University of Melbourne (Emeritus)

Norris M. Haynes, Southern Connecticut State University

JoAnn Hsueh, MDRC

Kristen Huff, Curriculum Associates

Diana Hughes, Relay Graduate School of Education

Gerunda B. Hughes, Howard University (Emeritus)

Neal Kingston, University of Kansas

Geoffrey T. LaFlair, Duolingo, Inc.

Carol D. Lee, Northwestern University (Emeritus)

Paul G. LeMahieu, Carnegie Foundation for the Advancement of Teaching; University of Hawai'i, Mānoa

Richard M. Lerner, Tufts University

Lei Liu, Educational Testing Service

Ou Lydia Liu, Educational Testing Service

Silvia Lovato, PBS KIDS

Temple S. Lovelace, Assessment for Good, Advanced Education Research and Development Fund (AERDF)

Susan Lyons, Lyons Assessment Consulting

Scott F. Marion, National Center for the Improvement of Educational Assessment

Kimberly McIntee, University of Massachusetts Amherst

Maxine McKinney de Royston, Erikson Institute

Elizabeth Mokyr Horner, Gates Foundation

Orrin T. Murray, The Wallis Research Group

Na'ilah Suad Nasir, Spencer Foundation

Michelle Odemwingie, The Achievement Network, Ltd.

Maria Elena Oliveri, Purdue University

Saskia Op den Bosch, RevX

V. Elizabeth Owen, Age of Learning

Trevor Packer, College Board

Roy Pea, Stanford University

James W. Pellegrino, University of Illinois Chicago

Mario Piacentini, Organisation for Economic Cooperation and Development (OECD)

Mya Poe, Northeastern University

Ximena A. Portilla, MDRC

Elizabeth J. K. H. Redman, University of California, Los Angeles, Center for Research on Evaluation, Standards, & Student Testing (CRESST), School of Education & Information Studies (SE&IS)

Jeremy D. Roberts, PBS KIDS

Mary-Celeste Schreuder, The Achievement Network, Ltd.

David Sherer, Carnegie Foundation for the Advancement of Teaching

Stephen G. Sireci, University of Massachusetts Amherst, Center for Educational Assessment

Erica Snow, Roblox

Rebecca A. Stone-Danahy, College Board

Rebecca Sutherland, Reading Reimagined, Advanced Education Research and Development Fund (AERDF)

Natalya Tabony, College Board

Carrie Townley-Flores, Rapid Online Assessment of Reading (ROAR), Stanford University

Eric M. Tucker, The Study Group

Alina A. von Davier, Duolingo, Inc.

Kevin Yancey, Duolingo, Inc.

Jessica W. Younger, PBS KIDS

Constance Yowell, Northeastern University

Biographical Statements

Sergio Araneda, Ph.D., is a research scientist specializing in educational measurement, psychometrics, and test security. He earned his doctorate in Research, Educational Measurement, and Psychometrics from the University of Massachusetts Amherst, following completion of his undergraduate studies in Mathematical Civil Engineering at the Universidad de Chile. Dr. Araneda currently works at Caveon, where he investigates how large language models can be integrated with test security innovations such as SmartItems™, contributing to research, publications, and conference presentations. He previously served as an associate psychometrician at the College Board, focusing on item parameter drift and automated essay scoring for the SAT, and as a research assistant at DEMRE, Universidad de Chile, evaluating policies in university admissions. His earlier professional experience also includes roles in finance as a quantitative analyst and consultant, providing him with a strong technical and analytical background. His academic and professional contributions span peer-reviewed publications, white papers, newspaper columns, and numerous presentations at international conferences, including NCME, ITC, and ATP. He also serves as Vice-Coordinator of FEVED, a professional forum advocating for best practices in educational assessment in Chile.

Eleanor Armour-Thomas, Ed.D., is Professor Emerita at Queens College, CUNY, where she served in the Department of Secondary Education from 1987 to 2024, including 22 years (2000–2022) as Department Chair. She specialized in Educational Psychology, teaching pre-service and in-service teachers, and served as Principal Investigator and Co-Principal Investigator for programs aimed at enhancing mathematics teacher preparation and professional development in science education. Her books, journal articles, oral addresses, and reports focus on teacher and student cognition, metacognition, learning, and assessment. Additionally, she has evaluated educational programs designed to improve learning and academic achievement for students from low socio-economic backgrounds and has consulted on teaching, learning, and assessment in K-16 education.

Aneesha Badrinarayan is a Principal Consultant at Education First, where she partners with state and district leaders, assessment developers, and policymakers to design coherent systems of teaching, learning, and assessment. She brings decades of expertise in assessment design, STEM education, policy, and product development, helping organizations and leaders create and implement instructionally relevant assessment systems. At Education First, Aneesha leads projects on innovative assessment and accountability design, equitable assessment, strategic planning, and artificial intelligence. Previously, Aneesha directed assessment work at the Learning Policy Institute, leading innovations across 15 states, shaping the 2028 NAEP Science Framework, and guiding federal policy on learning-first assessments. A behavioral neuroscientist by training, she holds degrees from Cornell University and the University of Michigan.

Eva L. Baker is a Distinguished Professor at UCLA and founding Director of the Center for Research on Evaluation, Standards and Student Testing, (CRESST). She is widely published in the areas of learning-based assessments, technology, and policy. She served as Chair of the Board on Testing and Assessment, National Research Council, and Co-Chair of the 1999 Standards for Educational and Psychological Testing. Baker served as president of the World Education Research Association (WERA) and was president of the American Educational Research Association (AERA). A member of the National Academy of Education, she received AERA's Robert L. Linn Lecture and the E. F. Lindquist Award.

Dr. Hee Jin Bang, Vice President of Efficacy Research & Evaluation at Age of Learning, Inc., leads research initiatives evaluating the effectiveness of educational technology products. In her current role, she oversees research studies examining the impact of adaptive learning technologies on student achievement across diverse populations and educational settings. Her recent publications offer compelling evidence for the effectiveness of digital learning platforms, demonstrating significant learning gains in language acquisition, early mathematics, and reading skills. Currently, as co-principal investigator on a $3.5 million Institute of Education Sciences-funded study, she continues to shape more effective educational technology solutions by investigating how personalized game-based learning supports teaching and learning in classrooms. Prior to joining Age of Learning, she held research leadership positions at Classroom, Inc., Amplify Education, and National Writing Project, where she evaluated digital curricula, assessments, and teacher professional development programs. She holds a Ph.D. from NYU in Teaching & Learning, an M.Ed. in Human Development and Psychology from Harvard University, and a B.A. (Honors) in Linguistics and French from Oxford University.

Héfer Bembenutty, Ph.D., is dedicated to advancing the field of educational psychology through his role as a professor at Queens College, The City University of New York. His academic journey led him to earn a Ph.D. in educational psychology from the same institution. Dr. Bembenutty's research focuses on the self-regulation of learning among high school and college students, as well as teachers. He explores various aspects such as assessment, homework self-regulation, self-efficacy beliefs, culturally self-regulated pedagogy, and academic delay of gratification. His teaching portfolio includes undergraduate and graduate courses on educational psychology, cognition, instruction and technology, human development and learning, assessment and measurement, and classroom management. Additionally, he investigates the impact of demographic factors like gender and ethnicity on students' ability to prioritize long-term goals over immediate rewards. He is an accomplished author and editor, contributing to several books and peer-reviewed journals. His work integrates contemporary theories with practical applications to enhance self-regulated learning in educational environments.

Randy E. Bennett holds the Norman O. Frederiksen Chair in Assessment Innovation in the ETS Research Institute. His recent work centers on personalized assessments and, relatedly, assessments that are "born socioculturally responsive." From 1999–2005 he directed the National Assessment of Educational Progress (NAEP) Technology-Based Assessment project, which included the first administration of computer-based performance assessments to nationally representative samples of U.S. school students and the first use of logfile data in such samples to measure problem-solving processes. From 2007–2016, he directed the CBAL research initiative (Cognitively Based Assessment of, for, and as Learning), which created theory-based summative and formative assessment to model good teaching and learning practice. He is a past president of the International Association for Educational Assessment and of the National Council on Measurement in Education (NCME). He is a fellow of the American Educational Research Association (AERA) and an elected member of the National Academy of Education, as well as recipient of the NCME Bradley Hanson Contributions to Educational Measurement Award, the Teachers College Columbia University Distinguished Alumni Award, the AERA E. F. Lindquist Award, and the AERA Cognition and Assessment SIG Award for Outstanding Contribution to Research in Cognition and Assessment.

Dr. Anastasia Betts is a leading expert in education and learning sciences innovation. As Executive Director of Learnology Labs, a collaborative think tank, she leads cutting-edge research on AI-enabled learning systems with a focus on transforming early childhood. Dr. Betts previously led the curriculum research, design, and production of digital learning products for early learning at Age of Learning, where her pioneering work in adaptive learning systems resulted in her inclusion on three U.S. patents. Currently, Dr. Betts spearheads the development of PAL (Personal Assistant for Learning), an AI-driven system that exemplifies distributed cognition principles to empower parents and teachers in supporting early math development. Dr. Betts holds a Ph.D. in Curriculum, Instruction, & the Science of Learning from the University at Buffalo, SUNY. Her research and publications focus on leveraging learning sciences and AI to create more equitable, personalized educational experiences. She is editor of the *Handbook of Research for Innovative Approaches to Early Childhood Education and Kindergarten Readiness* and has authored numerous papers on adaptive learning and human-AI partnerships in education. Dr. Betts was selected as a Harvard Women in Educational Leadership Fellow and was twice nominated for the American Educational Research Association (AERA) Karen King Future Leader Award.

Dr. Mary K. Boudreaux is an Associate Professor and Coordinator of the Doctoral Program in Educational Leadership & Policy Studies at Southern Connecticut State University. With a distinguished career spanning K-12 and higher education, she has served as a curriculum director, educational specialist, consultant, and university faculty member. Dr. Boudreaux specializes in improving school culture and climate, enhancing leadership practices, and promoting equity-focused practices and assessment strategies. As an educator and scholar, Dr. Boudreaux has designed and taught graduate and doctoral courses in organizational leadership, research methods, curriculum development, assessment, and change leadership. Her work prepares aspiring and practicing educational leaders to address systemic challenges through data-driven decision-making and evidence-based assessment practices. A prolific researcher, she has published numerous peer-reviewed articles, book chapters, and conference presentations on multicultural awareness and leadership, as well as fostering inclusive and equitable learning environments. Dr. Boudreaux's commitment to continuous improvement in education is reflected in her leadership roles as Co-Chair of the University Standards and Assessment Review Committee and a member of the University Graduate Council. These positions allow her to shape institutional assessment practices, ensuring academic programs achieve and maintain high-quality performance standards. Holding doctoral degrees in Educational Leadership and Innovation and Curriculum & Instruction, alongside certifications in higher education leadership, instructional design, and academic advising, Dr. Boudreaux remains dedicated to enhancing educational excellence and shaping future generations of scholars and practitioners.

Susan M. Brookhart, Ph.D., is Professor Emerita in the School of Education at Duquesne University and an independent educational consultant. She was the 2007-2009 Editor of Educational Measurement: Issues and Practice and is currently an Associate Editor of Applied Measurement in Education. She is the author or co-author of over 100 articles, chapters, and books on classroom assessment, teacher professional development, and evaluation. She was named the 2014 Jason Millman Scholar by the Consortium for Research on Educational Assessment and Teaching Effectiveness (CREATE) and was the recipient of the 2015 Samuel J. Messick Memorial Lecture Award from ETS/TOEFL. Dr. Brookhart's research interests include the role of both formative and summative classroom assessment in student motivation and achievement, the connection between classroom assessment and large-scale assessment, and grading. Dr. Brookhart received her Ph.D. in Educational Research and Evaluation from The Ohio State University, after teaching in both elementary and middle schools.

Dr. Carol Bonilla Bowman is an Associate Professor of Education at Ramapo College of New Jersey, where she also serves as a program director. Her research and publications focus on portfolios as both assessment and learning tools. Her recent work focuses on contemplative education. She holds a doctoral degree in applied linguistics and bilingual education from Teachers College, Columbia.

Dr. Sean P. "Jack" Buckley is Vice President of People at Roblox, where he oversees several teams including People (HR) and People Science and Analytics. He was previously President and Chief Scientist at Imbellus, Senior Vice President at the American Institutes for Research (AIR), and Senior Vice President of Research at The College Board. He also served as Commissioner of the U.S. Department of Education's National Center for Education Statistics (NCES) and as an Associate Professor at New York University, and an Assistant Professor at Boston College. He began his career as a surface warfare officer and nuclear reactor engineer in the U.S. Navy and has also worked in intelligence analysis. He holds an M.A. and Ph.D. in Political Science from Stony Brook University and an A.B. in Government from Harvard University.

Jill Burstein is Principal Assessment Scientist at Duolingo, leading validity and efficacy research for the Duolingo English Test – Duolingo's English language proficiency test. Her career has been motivated by social impact, working on AI-driven, education technology to enhance equity and access for learners and test takers. Her research lies at the intersection of artificial intelligence and natural language processing, educational measurement, equity in education, learning analytics, and linguistics. Dr. Burstein pioneered the first automated writing evaluation system used in large-scale, high-stakes assessment, as well as early commercial online writing instruction tools. She holds numerous patents for this work, and has published extensively in the field of AI in education, including topics in automated writing evaluation, digital assessment, responsible AI, and writing analytics. Her recent work focuses on responsible AI for digital assessment, and wrote the Duolingo English Test Responsible AI Standards, the first standards for an assessment program. Additionally, she is a co-founder of SIG EDU, an ACL Special Interest Group on Building Educational Applications. Dr. Burstein holds a Ph.D. in Linguistics from the Graduate Center, City University of New York.

Pamela Cantor, M.D., is a child and adolescent psychiatrist and the Founder and CEO of The Human Potential L.A.B., whose mission is to leverage scientific knowledge and technologies to transform what people understand and what institutions do to unlock human potential in each and every individual. Dr. Cantor is an author of Whole-Child Development, Learning and Thriving: A Dynamic Systems Approach (Cambridge University Press) and The Science of Learning and Development (Routledge). She founded the nonprofit organization Turnaround for Children (now the Center for Whole-Child Education at Arizona State University), is a Governing Partner of the Science of Learning and Development Alliance, and a strategic science advisor to the Carnegie Foundation for the Advancement of Teaching, the American Association of School Superintendents, and Learning Heroes. Dr. Cantor received an M.D. from Cornell University, a B.A. from Sarah Lawrence College, served as an Assistant Clinical Professor of Child Psychiatry at Yale School of Medicine, and was a Visiting Scholar at the Harvard Graduate School of Education.

Dr. Jennifer Charlot is co-founder of RevX, where she serves as Head of Programming. She leads the implementation of RevX's assessment system, ensuring data collection is integrated into daily instruction and shaping our systems for using real-time insights to refine teaching practice. As Managing Partner at Transcend, she directed early-stage school design projects, spreading science-driven innovation nationwide. A serial entrepreneur, Dr. Charlot spearheaded career and technical education programs for disconnected youth in NYC and served as Director of Implementation at Character Lab, translating research into practical classroom strategies. She holds a Doctorate in Education Leadership from Harvard's Graduate School of Education, a Master of Science in Social Administration from Columbia University, and a Bachelor of Arts from Boston College. Dr. Charlot is dedicated to reimagining educational systems through innovative design, actionable strategies, and data-driven practice—empowering young people to emerge as changemakers in their communities and beyond.

Gregory K. W. K. Chung, Ph.D. is the Associate Director for Technology and Research Innovation. Dr. Chung has extensive experience with the use of technology for learning and assessment. He has led projects related to game-based learning or game-based assessments involving pre-school students to adults in formal and informal settings with a focus on STEM topics (e.g., math, physics, engineering, programming) as well as social-emotional learning. His research involves small-scale exploratory studies to multi-district, multi-state RCT. He has conducted instructional technology R&D for IES, NSF, Office of Naval Research, PBS KIDS, Bill and Melinda Gates Foundation, Caplan Foundation for Early Childhood, and numerous other foundations and commercial entities.

Paul Cobb is Professor Emeritus at Vanderbilt University. His work focuses on improving the quality of mathematics teaching and student learning on a large scale. He is currently involved in a project that is developing practical measures of key aspects of high quality mathematics and investigating their use as levers for and measures of instructional improvement. He received Hans Freudenthal Medal for cumulative research program over the prior ten years from the International Commission on Mathematics Instruction (ICMI) in 2005, and the Silver Scribner Award from American Educational Research Association in 2010 for research over the past ten years that contributes to our understanding of learning and instruction.

Kimberly Cockrell is an experienced educator, administrator, and leader committed to instructional excellence, leadership development, and equity in education. With over two decades of experience in school leadership, professional learning, and strategic partnerships, she has worked to transform assessment and instructional practices to better support educators and students. At Achievement Network (ANet), Kimberly directs communications and stakeholder engagement, shaping public discourse around instructional coherence, data-driven decision-making, and student success. Kimberly's career spans charter, public, and independent schools, where she has designed professional development programs, led data-driven instructional strategies, and championed equitable learning environments. A lifelong learner and consultant, she continues to support educators in strengthening school leadership, assessment literacy, and instructional coherence.

Kelly Corrado is the Director of Game Tooling and Analytics Products for PBS KIDS. Corrado is committed to leveraging technology to enrich early childhood education through the delivery of high-impact products and experiences at scale for children aged 2-8 and the grownups who support them in school and in life. Corrado is a results-driven product leader with success leading cross-functional teams, optimizing digital ecosystems, and driving strategic initiatives that enhance accessibility, performance, and engagement. With a focus on game development and analytics platforms, Corrado influences business growth and user experience through data insights and innovation.

Danielle Crabtree, **M.Ed.**, is a doctoral student in the Research, Educational Measurement, and Psychometrics program at the University of Massachusetts Amherst. She holds dual master's degrees in Educational Administration and Secondary Education, and bachelor's degrees in Mathematics and Biochemistry & Molecular Biology, giving her a strong interdisciplinary foundation. Her research examines educational equity, teacher professional learning, and technology-enhanced instruction. She focuses on developing new methods to capture complex, hidden aspects of teaching and learning, broadening how assessment can inform both research and practice. As a Graduate Research Assistant, Danielle has contributed to WearableLearning, a game-based platform integrating embodied learning and computational thinking in mathematics led by Professor Ivon Arroyo, and EMPOWER, a research-practice partnership exploring the development of teacher educators' critical consciousness in science classrooms led by Associate Professor Enrique Suárez. She has co-authored multiple peer-reviewed conference proceedings, including a 2024 paper nominated for Best Design Paper at the International Conference of the Learning Sciences. An experienced educator and administrator, Danielle has served as a classroom teacher, assistant principal, practicum supervisor, and university instructor. She holds licensure as both a secondary teacher and PreK–12 principal. Passionate about advancing educational equity and innovation, she works to bridge research and practice to strengthen teacher development and improve outcomes for both teachers and students.

Linda Darling-Hammond is the Charles E. Ducommun Professor of Education, Emeritus, at Stanford University and founding president of the Learning Policy Institute, where she leads research and policy initiatives focused on educational equity, teacher quality, and effective school reform. A nationally renowned scholar, she has authored more than 30 books and hundreds of publications on teaching, learning, and education policy. Darling-Hammond's career has centered on advancing evidence-based policies that improve access to high-quality learning opportunities for all students. She served as chair of the California State Board of Education from 2019 to 2023, where she guided the state's efforts to strengthen curriculum, assessments, and teacher preparation. Earlier, she directed the Stanford Center for Opportunity Policy in Education and the National Commission on Teaching and America's Future, influencing reforms in teacher development and accountability systems across the U.S. Recognized as one of the most influential voices in education, she has advised federal and state leaders on issues ranging from school funding to equitable assessment design. Darling-Hammond continues to champion the creation of schools that support deep learning, social-emotional growth, and equitable outcomes for every child.

Jacqueline Darvin, Ph.D., is a Program Director and Professor of Literacy Education at Queens College of the City University of New York (CUNY). In addition to a BA in Psychology and doctorate in Literacy Studies, she has master's degrees in educational leadership and secondary education and credentials as a New York State School District Leader. Before becoming a professor at Queens College, Dr. Darvin taught middle and high school Title One reading, Special Education, and English for twelve years. In 2015, she published a book with Teachers College Press titled Teaching the Tough Issues: Problem-Solving from Multiple Perspectives in Middle and High School Humanities Classes. She was the recipient of the Long Island Educator of the Month Award, featured in a cover story of New York Teacher, the official publication of the New York State United Teachers' Union, and a recipient of the Queens College Presidential Award for Innovative Teaching. She is a workshop provider for Nassau and Easter Suffolk BOCES and provides consulting and professional development to schools and teachers throughout the New York metropolitan area. Her presentations include local, regional, national and international conferences on topics related to literacy teaching and learning.

Girlie C. Delacruz is Associate Vice Chancellor for Teaching and Learning at Northeastern University, where she oversees experiential learning programs in undergraduate research, service learning, and community and civic engagement, as well as student support through fellowships advising and peer tutoring. With over two decades of experience spanning research and applied practice, she has led initiatives to expand equitable access to education, including as Chief Learning Officer for LRNG at Southern New Hampshire University and as a researcher at UCLA developing technology-enhanced assessments for military and educational contexts. Her scholarship and leadership have been recognized through awards such as Northeastern's 2025 Staff Excellence Award for Mentorship and the APA Military Psychology Research Award, as well as fellowships from the MacArthur Foundation and ETS. She also serves on national grant review panels and has published widely on learning, assessment design, and the role of technology in advancing equity.

Clarissa Deverel-Rico, Ph.D., is a postdoctoral researcher at BSCS Science Learning. A former middle school science teacher, Clarissa transitioned into a career driven by creating better science learning experiences for students. She studies innovative approaches for how classroom assessment can support a vision of science education that prioritizes epistemic justice, care, and student experience. Current research aims include studying the extent to which currently available classroom assessments support equitable opportunities to learn, developing assessments for broad use in high school biology, investigating the efficacy of locally-adapted high-quality curricular materials, and partnering with teachers around creating spaces to learn directly from students and families for how classroom assessment can be spaces that sustain students' interests and identities.

Dr. Kristen Eignor DiCerbo is the Chief Learning Officer at Khan Academy, a nonprofit dedicated to providing a free world class education to anyone, anywhere. In this role, she is responsible for the research-based teaching and learning strategy for Khan Academy's offerings. She leads the content, assessment, design, product management, and community support teams. Time magazine named her one of the top 100 people influencing the future of AI in 2024. Dr. DiCerbo's work has consistently been focused on embedding what we know from education research about how people learn into digital learning experiences. Prior to her role at Khan Academy, she was Vice-President of Learning Research and Design at Pearson, served as a research scientist supporting the Cisco Networking Academies, and worked as a school psychologist in an Arizona school district. Kristen received her Bachelor's degree from Hamilton College and Master's degree and Ph.D. in Educational Psychology at Arizona State University.

Ravit Dotan, Ph.D., is a renowned tech ethicist specializing in artificial intelligence (AI) and data technologies. She aids tech companies, investors, and procurement teams in developing and implementing responsible AI strategies, conducts research on these topics and creates resources. Dr. Dotan was recognized as one of the 100 Brilliant Women in AI Ethics for 2023 and has received accolades such as the 2022 "Distinguished Paper" Award from the FAccT conference. Her views are frequently featured in prominent publications like the *New York Times, The Financial Times,* AP News, and TechCrunch. Dr. Dotan holds a Ph.D. in Philosophy from UC Berkeley and has extensive experience in AI ethics research, teaching, and advocacy for diversity and inclusion in academia. You can find Dr. Dotan's resources on her AI Ethics Treasure Chest and LinkedIn page.

Kerrie A. Douglas, Ph.D., is an Associate Professor of Engineering Education at Purdue University and Co-Director of SCALE, a large Department of Defense funded workforce development project in secure microelectronics. In that role, she leads the education and workforce development across 33 universities in the U.S. She is passionate about modernizing engineering education and preparing learners for their professional work. Her research is focused on improving methods of evaluation and assessment in engineering learning contexts. She works on assessment problems in engineering education, such as considerations for fairness, how to assess complex engineering competencies, and aligning assessment to emerging workforce needs. She has been Primary Investigator or Co-PI on more than $100 million of external research awards. In 2020, she received an NSF RAPID award to study engineering instructional decisions and how students were supported during the time of emergency remote instruction due to the COVID-19 pandemic. In 2021, she received the NSF CAREER award to study improving the fairness of assessment in engineering classrooms. She has published over 100 peer-reviewed journal and conference papers.

Dr. Kadriye Ercikan is the Senior Vice President of Global Research at the Educational Testing Service (ETS), President and CEO of ETS Canada Inc., and Professor Emerita at the University of British Columbia. In these leadership roles, she directs foundational and applied research. Her research focuses on validity and fairness issues and sociocultural context of assessment. Her recent research includes validity and fairness issues in innovative digital assessments, including using response process data, AI applications, and adaptivity. Ercikan is the President and a Fellow of the International Academy of Education (IAE), President of the International Test Commission (ITC), and President-Elect of the National Council on Measurement in Education (NCME). Her research has resulted in six books, four special issues of refereed journals and over 150 publications. She was awarded the AERA Division D Significant Contributions to Educational Measurement and Research Methodology recognition for another co-edited volume, Generalizing from Educational Research: Beyond Qualitative and Quantitative Polarization, and received an Early Career Award from the University of British Columbia. Ercikan is currently serving as the NCME Book Series Editor (2021-2026).

David S. Escoffery is a Director in the Graduate and Professional Education area at Educational Testing Service. He joined ETS in 2006 after teaching theatre history at the university level for five years. His academic areas of specialization include theatre history and literature, English language and literature, pedagogical theory, and cultural studies. He applies his experience to the development of examinations that measure knowledge of critical thinking, writing, and analytical reasoning. In addition to AP Art and Design, he has worked on a wide variety of assessment programs, including GRE, Praxis, and SAT. He has published numerous articles in journals such as Applied Measurement in Education and served as the editor for the 2006 McFarland collection How Real Is Reality TV? He earned his Ph.D. and M.A. in theatre history, literature, and criticism from the University of Pittsburgh, and his A.B. in English from Princeton University.

Carla M. Evans is a Senior Associate at the National Center for the Improvement of Educational Assessment, where she leads efforts to develop and implement balanced assessment and accountability systems for states, bridging the classroom and policymaking levels. Carla's work spans systemwide assessment reviews, assessment literacy initiatives, performance-based assessment design, and aligning accountability systems with educational values. Her research emphasis lies in culturally responsive assessment, competency-based education, AI in classroom assessment, and instructionally useful assessment.

Howard T. Everson is a Professor of Educational Psychology (by courtesy) at the Graduate School, City University of New York. He is the former Director of the Center for Advanced Study in Education at the Graduate School, City University of New York. His research and scholarly interests focus on the intersection of cognition, technology and assessment. He has published widely and has contributed to developments in educational psychology, psychometrics, quantitative methods, and program evaluation. Professor Everson's measurement expertise is in the areas of evidence-centered design, item response theory, differential item functioning, learning analytics and cognitive diagnostic measurement models. Dr. Everson also served as the Executive Director of the NAEP Educational Statistics Services Institute at the American Institutes for Research, and was the Vice President and Chief Research Scientist at the College Board. Dr. Everson is a Psychometric Fellow at the Educational Testing Service, and an elected Fellow of both the American Educational Research Association and the American Psychological Association, and a charter member of the Association for Psychological Science. Dr. Everson is the former editor of the National Council of Measurement in Education's journal, Educational Measurement: Issues and Practice.

Cosimo Felline, Ph.D., is the Director of Data Science and Analytics at PBS KIDS. With a background in theoretical nuclear physics, he earned his doctorate before transitioning from academia to the tech industry. Beginning his career as a web developer, software engineer, and manager, Felline developed a strong foundation in software development and web technologies. More recently, he has shifted his focus to data science and engineering, where he applies his expertise to building scalable data solutions. Passionate about data literacy and democratization, he is committed to breaking down barriers to data access and enabling actionable insights. He enjoys playing the piano, watching horror movies, and petting his dogs.

Kate Felsen is the Chief Communications Officer of The Human Potential L.A.B. and President of Up Up Communications LLC, with clients focused on transforming education and supporting healthy youth development. Kate had a distinguished career at ABC News. As Foreign Editor for the flagship evening news broadcast, she covered breaking and feature stories around the globe, winning 11 Emmy Awards. Kate earned an M.A. in American foreign policy and international economics from Johns Hopkins and a B.A., *magna cum laude* in history and literature from Harvard. She garnered first-team All-American and Ivy League "Player of the Year" honors in lacrosse, captained the field hockey team and enjoys coaching a club lacrosse team for middle school girls in New York City. She serves as Chair of the Board of USA Climbing and Feed the Frontlines NYC.

Tianying Feng is a Ph.D. candidate in the Education – Advanced Quantitative Methods program at UCLA and a research assistant at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), SEIS Building, Los Angeles, CA 90095-1522; tfeng0315@ucla.edu. Her primary research interests include technology-based measurement and learning, psychometrics, process modeling, and statistical computing.

Natalie Foster is an Analyst in the Programme for International Student Assessment (PISA) at the Organisation for Economic Cooperation and Development (OECD). Her work mainly focuses on the design and development of innovative assessments of 21st century competences included in each PISA cycle, working closely with measurement and test development experts, as well as various other PISA research and development projects. She is the lead author of the PISA 2022 Creative Thinking and PISA 2025 Learning in the Digital World assessment frameworks, co-editor of the publication Innovating Assessments to Measure and Support Complex Skills, and the lead author of the PISA 2022 Results (Volume III): Creative Minds, Creative Schools report. She has also worked in the OECD Centre for Educational Research and Innovation on the Smart Data and Digital Technologies in Education project, where she contributed to the OECD Digital Education Outlook 2023. Before joining PISA, she worked at the OECD Development Centre and European Commission.

James Paul Gee is a Regents Professor Emeritus at Arizona State University. He was, in his career, a professor at six universities. He is an elected member of the National Academy of Education. He received his Ph.D. in linguistics in 1975 from Stanford University and initially worked on syntactic theory and the philosophy of language, later becoming interested in a variety of other areas, including psycholinguistics, discourse analysis, sociolinguistics, literacy studies, learning theory, and video games. His books include Sociolinguistics and Literacies; The Social Mind; An Introduction to Discourse Analysis; Situated Language and Literacies; What Video Games Have to Teach Us About Literacy and Learning; The Anti-Education Era; and What is a Human? His current work is about the paradox that while we say "humans learn from experience" and experience is composed of sensory interactions with the world, we hear precious little about sensation in educational research.

Sheryl L. Gómez, serves as the Chief Financial and Operating Officer for the Study Group, where she leads strategy, finance, and operations to advance equity, innovation, and impact in education. She is a results-driven finance and operations executive across the public, private, and social sectors. She has served as the CFO for Brooklyn Laboratory Charter Schools, CFO and COO of Friends of Brooklyn LAB, CFO and COO of Equity By Design, a Financial Manager at Charter School Business Management, and a Financial Manager at FOREsight Financial Services for Good. Her experience includes managing clients' accounts, maintaining accurate records of financial transactions, financial reports, monthly close reviews, financial audits, and year-end processes. She has expertise in organizational growth, resource development, financial strategy, and public-private partnerships. She has managed multimillion-dollar budgets, secured over $150M in facilities financing, and overseen grants from major funders.

Edmund W. Gordon is the John M. Musser Professor of Psychology, Emeritus at Yale University; Richard March Hoe Professor, Emeritus of Psychology and Education, at Teachers College, Columbia University; Director Emeritus of the Edmund W. Gordon Institute for Advanced Study, at Teachers College, Columbia University; and Honorary President of the American Educational Research Association. Gordon's distinguished career spans professional practice and scholarly life as a minister, clinical and counseling psychologist, research scientist, author, editor, and professor. He earned his B.S. in Zoology and B.D. at Howard University, an M.A. in Social Psychology from American University, and an Ed.D. in Child Development and Guidance from Teachers College, Columbia University. He received the AERA Relating Research to Practice Award (2010), the John Hope Franklin Award (2011), and the Harold W. McGraw, Jr. Prize in Education (2024). He is widely recognized for his work on the Head Start program, the achievement gap, supplementary education, the affirmative development of academic ability, and Assessment in the Service of Learning. Author of more than 400 articles and 25 books, Gordon has been named one of America's most prolific and thoughtful scholars. He was married to Susan Gitt Gordon for 75 years and together had four children.

Sunil Gunderia, is Chief Innovation Officer at Age of Learning, the company behind ABCmouse, an early learning program trusted by the parents of 50 million children. He co-invented the AI-based personalized mastery learning system powering My Math Academy and My Reading Academy, game-based programs whose effectiveness has been validated by 28 ESSA-aligned studies. Research finds over 90 percent of teachers want these programs for their impact on learning and on students' confidence and interest in reading and math. Sunil is Vice Chair of the EdSAFE AI Industry Council and Advisor to National AI Literacy Day and the Center for Outcome-Based Contracting. He also serves on the boards of InnovateEDU and the Children's Institute, which provides Head Start and mental health services to more than 30,000 children and families. Previously, he worked for The Walt Disney Company, where he ran the global mobile games business after starting it in Europe.

Laura S. Hamilton is a senior associate at the National Center for the Improvement of Educational Assessment, where she collaborates with states, districts, and nonprofit organizations on the design and implementation of assessment policies and practices. She is especially interested in supporting the development and implementation of large-scale and classroom assessment systems that measure students' civic readiness, and she is co-editing a volume on assessing civic learning and engagement. Her previous roles include senior director at American Institutes for Research, associate vice president in the Research and Measurement Sciences area at ETS, distinguished chair in learning and assessment at RAND, and co-director of RAND's nationally representative educator survey panels. Hamilton regularly serves on expert committees and panels including the Joint Committee to revise the AERA/APA/NCME Standards for Educational and Psychological Testing, multiple National Academies of Sciences, Engineering, and Medicine committees, and technical advisory committees for state assessment programs. She's also held editorial roles with several journals. She is a fellow of the American Educational Research Association and received the Joseph A. Zins Distinguished Scholar Award for Social and Emotional Learning Research. Hamilton earned a Ph.D. in educational psychology and an M.S. in statistics from Stanford University.

Emily C. Hanno is a Senior Research Associate at MDRC where she is Project Director and co-Principal Investigator of the Measures for Early Success Initiative. Hanno's research, which is grounded in her experiences as a Head Start teacher and instructional coach, focuses on understanding how early education and care innovations, programs, and policies can support children, families, and communities.

John Hattie is Emeritus Laureate Professor at the Melbourne Graduate School of Education at the University of Melbourne, Chief Academic Advisor for Corwin, i-Ready Technical Advisor, and co-director of the Hattie Family Foundation. His career was as a measurement and statistics researcher and teacher, and his more recent research, better known as Visible Learning, is a culmination of nearly 30 years synthesizing more than 2,500 meta-analyses comprising more than 140,000 studies involving over 300 million students around the world.

Dr. Norris M. Haynes is a Professor in the Educational Leadership Department at Southern Connecticut State University. He founded and directed the Center for Community and School Action Research (CCSAR) and served as Chairperson of the Counseling and School Psychology Department. Dr. Haynes is a Clinical faculty member at the Yale University School of Medicine Child Study Center and where he has been an Associate Professor and Director of Research for the Yale University Comer School Development Program. He earned his Ph.D. in Educational Psychology and an M.B.A. with a focus on health services administration from Howard University. Haynes is a licensed Psychologist, Fellow of the American Psychological Association, and Diplomate in the International Academy for Behavioral Medicine, Counseling, and Psychotherapy. His research interests include social-emotional learning, school climate, resilience, and academic achievement. Dr. Haynes has authored numerous articles, books, and evaluation reports. He is a founding leadership team member of the Collaborative for Academic and Social Learning (CASEL) and researcher with Social Emotional and Character Development (SECD). He has worked with educational and psychological entities to enhance school practices. Dr. Haynes has been involved in national research initiatives, including studies on youth violence, social and emotional learning, and the Harlem Children's Zone (HCZ) programs.

JoAnn Hsueh is currently Vice President of Program and Communications at the Foundation for Child Development and co-Principal Investigator and Senior Advisor for the Measures for Early Success Initiative. Trained as a developmental scientist, Hsueh has broad interests in studying the impact and implementation of social, economic, and educational policies and programs that influence family and child well-being.

Kristen Huff, M.Ed., Ed.D., currently serves as the Head of Measurement at Curriculum Associates, where she leads a team of assessment designers, psychometricians, and researchers in the development of online assessments integrated with personalized learning and teacher-led instruction. Prior to this role, she served as the Senior Fellow for the New York State Education Department as well as serving in leadership roles with several major assessment companies. Dr. Huff has deep expertise in k-12 large scale assessment, and has presented and published consistently in educational measurement conferences and publications for over 25 years. She served previously as a technical advisor for the 2026 NAEP Frameworks in Reading and Mathematics and as the inaugural Co-Chair of the NCME Task Force on Classroom Assessment 2016-2020. She was named as recipient of the 2021 Career Achievement Award from the Association of Test Publishers, and now serves as the NCME Representative to the Management Committee for the revision of the 2014 Joint Standards for Educational and Psychological Testing, published by AERA, APA, and NCME. Dr. Huff is first author of the forthcoming *Educational Measurement, 5th Edition* (Oxford University Press), and Designing and Developing Educational Assessments (Huff, Nichols, and Schneider).

Diana Hughes is Head of Product at Relay Graduate School of Education. She is an experienced practitioner of game design and personalized learning. As VP of Learning Science and Design at Age of Learning, Inc., Diana led the development of Age of Learning's science-backed, evidence-centered programs, My Math Academy, My Reading Academy, and My Reading Academy Español. With three patents in personalized learning technologies to her name, Diana is known for her innovative and effective contributions to digital education methodologies. Her work, underpinned by a profound commitment to student-centric design and efficacy, exemplifies her dedication to providing equitable, effective, and engaging learning experiences for children globally. Diana's past work includes an empathy game for children on the autism spectrum, a graphics-free game for blind and low-vision players, and soft skills training games for the United States Military. She holds an MFA in Game and Interactive Design from the University of Southern California and a BS in Multimedia from Bradley University.

Gerunda B. Hughes is Professor Emerita, Howard University. During her tenure at the University, Dr. Hughes served as Director of the Office of Institutional Assessment and Evaluation and Professor of Mathematics Education. As Director, she oversaw the collection and analyses of student learning and other institutional-level data. She also served as coordinator of secondary education programs and taught courses in mathematics, mathematics pedagogy, assessment and measurement, and research methodology. Dr. Hughes served as Principal Investigator of the "Classroom Assessment Project" at Howard University's Center for Research on the Education of Students Placed at Risk (CRESPAR). She was an inaugural member of the Board of Directors of the Howard University Middle School for Mathematics and Science. Dr. Hughes has served as Co-Editor-in-Chief of the Journal of Negro Education; Associate Editor of Review of Educational Research; and a member of the editorial boards of the American Educational Research Journal and the Mathematics Teaching-Research Journal. She currently serves on technical advisory committees for national, state, and professional testing and assessment organizations. Dr. Hughes earned a B.S. in mathematics from the University of Rhode Island, a M.A. in mathematics from the University of Maryland-College Park, and a Ph.D. in educational psychology from Howard University.

Neal Kingston, Ph.D., is University Distinguished Professor in the Department of Educational Psychology at the University of Kansas, Director of the Achievement and Assessment Institute (AAI), and Vice Provost for Jayhawk Global and Competency-Based Education. His research focuses on large-scale assessment, with particular emphasis on how it can better support student learning through the use of learning maps and diagnostic classification models. Current interests include games-based assessment, personalizing assessments to improve student engagement, and the creation of more agile test development approaches. Dr. Kingston has served as principal investigator or co-principal investigator for over 250 research grants. Of particular note was the Dynamic Learning Maps Alternate Assessment grant from the US Department of Education, which was at that time was the largest grant in KU history and which currently serves 23 state departments of education. Other important testing projects include the Kansas Assessment Program, Project Lead The Way, and Adaptive Reading Motivation Measures. He is known internationally for his work on large-scale assessment, formative assessment, and learning maps. He has served as a consultant or advisor for organizations such as the AT&T, College Board, Department of Defense Advisory Committee on Military Personnel Testing, Edvantia, General Equivalency Diploma (GED), Kaplan, King Fahd University of Petroleum and Minerals, Merrill Lynch, National Council on Disability, Qeyas (Saudi Arabian National Center for Assessment in Higher Education), the state of New Hampshire, the state of Utah, the U.S. Department of Education, and Western Governors University.

Geoffrey T. LaFlair is a Principal Assessment Scientist at Duolingo where he co-leads Assessment Research and Development for the Duolingo English Test. He holds an MA in TESOL from Central Michigan University and a Ph.D. in Applied Linguistics from Northern Arizona University. Prior to joining Duolingo, he was an Assistant Professor in the Department of Second Language Studies at the University of Hawai'i at Mānoa and the Director of Assessment in the Center for ESL at the University of Kentucky. His research interests are situated at the intersection of language assessment, psychometrics, and natural language processing, focusing on the application of research from these fields in researching and developing operational language assessments.

Carol D. Lee is the Edwina S. Tarry Professor Emeritus of Education in the School of Education and Social Policy and in African-American Studies at Northwestern University, and the President of the National Academy of Education. She is currently Chairman of the National Board of Education Sciences. She is a past president of the American Educational Research Association (AERA) and past president of the National Conference on Research in Language and Literacy. She is a member of the American Academy of Arts and Sciences and a fellow of the American Educational Research Association. She has won numerous awards and honors, including the McGraw Prize in Education. Her research addresses cultural supports for learning that include a broad ecological focus, integrating learning sciences and human development framing, with attention to language and literacy and African American youth. She is the author or co-editor of eleven books, monographs and special issues, including co-editing The Handbook of Cultural Foundations of Learning, and has published over 108 journal articles and book or handbook chapters in the field of education. She has also worked as an English Language Arts teacher and a primary grade teacher. She is a founder of four African-centered schools.

Paul G. LeMahieu is Senior Fellow at the Carnegie Foundation for the Advancement of Teaching and graduate faculty in education, University of Hawai'i at Mānoa. LeMahieu served as Superintendent of Education for the State of Hawai'i, serving 190,000 students. Prior to that, he was Undersecretary for Education Policy and Research for the State of Delaware. He has been President of the National Association of Test Directors and Vice President of the American Educational Research Association. He served on the National Academy of Sciences' Board on International Comparative Studies in Education, Mathematical Sciences Board, National Board on Testing Policy, and the National Board on Professional Teaching Standards. His professional interests focus on the adaptation of improvement science methodologies for application in networks in education. He is a co-author of the book *Learning to Improve: How America's Schools Can Get Better at Getting Better* (2015), and lead editor of the volume *Working to Improve: Seven Approaches to Improvement Science in Education* (2017). His most recent book is entitled *Measuring to Improve: Practical Measurement to Support Continuous Improvement in Education* (2025). Paul has a Ph.D. from the University of Pittsburgh, an M.Ed. from Harvard University, and an A.B. from Yale College.

Richard M. Lerner is the Bergstrom Chair in Applied Developmental Science and the Director of the Institute for Applied Research in Youth Development at Tufts University. He went from kindergarten through Ph.D. within the New York City public schools, completing his doctorate at the City University of New York in 1971 in developmental psychology. Lerner has more than 800 scholarly publications, including 90 authored or edited books. He was the founding editor of the Journal of Research on Adolescence and of Applied Developmental Science. He is currently the Editor of Review of General Psychology, the flagship journal of Division 1 of the American Psychological Association (APA). Lerner was a 1980-81 fellow at the Center for Advanced Study in the Behavioral Sciences and is a fellow of the American Association for the Advancement of Science, the APA, and the Association for Psychological Science (APS). He is the recipient of several awards for his career achievements: The SRA John P. Hill Memorial Award for Life-Time Outstanding work (2010); the APA Division 7 Urie Bronfenbrenner Award for Lifetime Contribution to Developmental Psychology in the Service of Science and Society (2013); the APA Gold Medal for Life Achievement in the Application of Psychology (2014); the APA Division 1 Ernest R. Hilgard Lifetime Achievement Award for distinguished career contributions to general psychology (2015); the ISSBD Award for the Applications of Behavioral Development Theory and Research (2016); the SRCD Distinguished Contributions to Public Policy and Practice in Child Development Award (2017); the APS James McKeen Cattell Fellow Award winner for lifetime outstanding contributions to applied psychological research (2020); and the SSHD Distinguished Lifetime Career Award (2021). Lerner served on the Board of Directors of the Military Child Education Coalition for 10 years and still serves on their Scientific Advisory Board. In February 2023, Pope Francis reappointed Lerner to a second five-year term as a Corresponding Member of the Pontifical Academy for Life.

Lei Liu is a Research Director leading the K–12 research team at ETS. She is also an Adjunct Professor at the University of Pennsylvania. Her research interests lie at the intersection of science learning and assessment, learning sciences, and educational technology. She has led multiple federal grants to develop transformative innovations for STEM learning, including topics on learning progressions, AI-supported assessment tools, and virtual labs. She has produced over 70 peer-reviewed publications. She is a member of the editorial board of Instructional Science and has served as a reviewer for multiple international conferences, journals, and NSF merit reviews. In addition to her lead role in research, Dr. Liu has also been a key contributor to support various operational works at ETS including the California State Assessment programs, and NAEP science and mathematics programs. She earned a Ph.D. in educational psychology with a focus on learning sciences and educational technology from Rutgers University.

Ou Lydia Liu, Associate Vice President of Research at ETS, is a globally recognized expert in assessment of critical skills and competencies in higher education and workforce. She has also managed large-scale grants awarded by government and private funding agencies in the U.S. and international countries including India, China, and Korea. Dr. Liu has authored and coauthored over 100 peer-reviewed journal articles, research reports, and book chapters in the fields of applied measurement, higher education, and science assessment. Her research appeared in Science, Nature Human Behavior, Educational Researcher, and other influential outlets. She delivered over 100 invited seminars and peer-reviewed conference presentations domestically and internationally. Dr. Liu was inducted as an AERA Fellow in 2023, and received the 2019 Robert Linn Memorial Lecture Award, and the 2011 National Council on Measurement in Education Jason Millman Promising Measurement Scholar Award in recognition of her original and extensive research in learning outcomes assessment in higher education and K-12 science assessment. Dr. Liu holds a doctorate in Quantitative Methods and Evaluation from the University of California, Berkeley.

Silvia Lovato is head of Learning & Research at PBS KIDS, where she leads the team responsible for PBS KIDS curriculum development, research and evaluation, and early childhood education strategy. Previously, she worked at PBS KIDS from 2000 to 2014 as a Content Manager and Senior Product Director, managing the production of interactive features for PBS KIDS digital platforms, especially games. A seasoned children's media professional and researcher who is passionate about how media can help kids learn, Silvia holds a Ph.D. in Media, Technology and Society from Northwestern University. Her dissertation, titled "Hey Google, Do Unicorns Exist?", explored how children use AI-based conversational agents such as the Google Assistant to seek answers to their many questions. She holds certificates in Cognitive Science and Management for Scientists and Engineers.

Dr. Temple S. Lovelace is the Executive Director of Assessment for Good (AFG), an inclusive R&D program supported by the Advanced Education Research and Development Fund (AERDF). AFG focuses on creating new assessment tools that explore how we recognize and maximize each student's potential as they leverage a unique set of skills to power their personal learning journey. In 2018, Temple launched a groundbreaking cooperative incubator in the School of Education at Duquesne University. There, she developed an innovative research and development methodology now being implemented by organizations across the United States. Her successful community-engaged programs—Youth Leading Change, Education Uncontained, and Girlhood Rising—have empowered educators and students to conduct localized R&D that bridges innovation and effective learning practices. Now, as a visiting scholar at the Gordon Institute for Advanced Study at Teachers College, Columbia University, Temple's research explores the role of context-capable assessment and learning so that we can understand the fullness of how learners explore their world and translate that to more modernized understandings of child development. A respected voice in educational innovation, Temple has published extensively on assessment design and student-centered learning approaches with the hope that educators, caregivers, and even learners themselves can co-create a future where all learners thrive.

Susan Lyons, Ph.D., works to transform traditional assessment systems to better serve the needs of students, educators, and the public. As the Principal Consultant at Lyons Assessment Consulting, Susan partners with innovators to advance theory and practice in educational measurement. Susan holds a bachelor's degree in Mathematics and Math Education from Boston University and served as a math educator before pursuing her graduate work. She received her master's and Ph.D. in Educational Psychology with a focus on Research, Evaluation, Measurement and Statistics from the University of Kansas. Susan is the co-founder of Women in Measurement, a nonprofit organization dedicated to advancing gender and racial equity in the field. Since its launch, she has served as the organization's Executive Director, ushering it through the start-up phase to its now prominent position as a fixture within the measurement community, offering support for more than a thousand women in our field.

Scott F. Marion, Ph.D., is a principal learning associate at the National Center for the Improvement of Educational Assessment. He is a national leader in conceptualizing and designing innovative and balanced assessment systems to support instructional and other critical uses. He has also led extensive work across the country to design and implement school accountability systems. Scott is an elected member of the National Academy of Education and is one of three measurement specialists on the National Assessment Governing Board, which oversees the National Assessment of Educational Progress. He coordinates and/or serves on 10 state or district technical advisory committees for assessment and accountability. He has served on multiple National Research Council committees, including those that provided guidance for next-generation science assessments, investigated the issues and challenges of incorporating value-added measures in educational accountability systems, and outlined best practices in state assessment systems. Scott is a co-author of the validity chapter in the 5th edition of Educational Measurement, a co-editor of the National Academy of Education's Reimagining Balanced Assessment, and a co-author of Instructionally Useful Assessment. He has published dozens of articles in peer-reviewed journals and edited volumes, and he regularly presents his work at the national conferences of the American Educational Research Association, National Council on Measurement in Education, and the Council of Chief State School Officers. Scott earned a Ph.D. from the University of Colorado Boulder with a concentration in measurement and evaluation.

Kimberly McIntee centers social (in)justice in developing equitable academic and assessment strategies and improving how results are created and shared. Her research examines testing procedures, assessment theories, and critiques of the harm curricula and assessments can cause individuals and society, with the goal of transforming traditional testing into meaningful practices that support teaching and learning. Growing up in a multiracial, multilingual environment pushed McIntee to constantly reflect on her identity and experiences across psychological, physical, and social dimensions. McIntee's earliest school memories involve navigating between worlds. This divide deepened when she and a few other minoritized peers were placed in classes where, despite attending predominantly Black schools, the majority of students became invisible in halls saturated with unfamiliar white faces. Such segregation often stemmed from curricula and assessments designed without accounting for diverse learners, particularly those least prepared by inequitable systems. Recognizing these hidden patterns of separation, McIntee advocates for schools where students' identities do not isolate them and where statistics do not dictate resources. She believes that through intentional research and just assessment design, academic and social spaces—long marked by inequity—can be reshaped into sites of empowerment.

Maxine McKinney de Royston is the Dean of Faculty at the Erikson Institute. Dr. McKinney de Royston's research and teaching examine how educators' political clarity can be reflected in their pedagogical practices in ways that support the intellectual thriving and holistic well-being of racially and economically minoritized learners. She is a co-editor, along with Na'ilah Suad Nasir, Erikson's Trustee Carol Lee, and Roy Pea, of the Handbook of the Cultural Foundations of Learning; free access: https://doi.org/10.4324/9780203774977. In addition to numerous peer-reviewed articles, chapters, and other publications and presentations, Dr. McKinney de Royston has served as Associate Editor of the American Educational Research Journal, Co-Chair of the Wallace Foundation Emerging Scholars Committee, and Advisor to the Wisconsin Department of Public Instruction, Family, Youth, & Community Advisory Council. She is a member of several professional learned societies, including the American Educational Research Association (AERA), the International Society of the Learning Sciences, the National Association for Multicultural Education, and the National Council of Black Studies.

Elizabeth Mokyr Horner is a Senior Program Officer at the Gates Foundation, which provided grant funding to support MDRC's Measures for Early Success Initiative. Dr. Mokyr Horner worked in partnership with MDRC to develop the approach to codesign described in this chapter. She has spent the last 15+ years across academic, non-profit, government, and foundation sectors supporting and evaluating evidence-based interventions designed to enhance educational outcomes, economic opportunity, and improved overall quality of life.

Orrin T. Murray, Ph.D., a learning scientist, is principal of the Wallis Research Group. Through Wallis Research Group, he has advised leading institutions, providing research, equity-driven program evaluations, and AI-based insights to shape social impact initiatives. He has been a workshop leader and mentor/coach, building evaluation skills and capacity in community-based organizations in Chicago and Cincinnati. As a Principal Researcher at the American Institutes for Research, he led national studies on education equity, civic education, AI-driven learning, and workforce development, ensuring that data-driven insights lead to real-world improvements. His thought leadership has shaped policy decisions, education strategies, and AI integration in learning, making him a trusted advisor to policymakers, school districts, and nonprofit organizations. At the University of Chicago's Urban Education Institute, he led a digital foundry responsible for designing and launching research-based tools to improve high school and college completion rates. Orrin's expertise extends into culturally responsive teaching, having contributed to "Culture in Our Classrooms," a documentary viewing guide on fostering belonging and inclusion in education. He is also a recognized voice in AI and education research, co-authoring "Principles to Guide Artificial Intelligence in Education Research", which outlines ethical considerations and bias mitigation in AI applications.

Na'ilah Suad Nasir is the sixth President of the Spencer Foundation, which funds education research nationally. Prior to joining Spencer, she held a faculty appointment in Education and African American Studies at the University of California, Berkeley where she also served as the chair of African American Studies, then later as the Vice Chancellor for Equity and Inclusion. Her scholarship focuses on race, culture, and learning, and how what we know about learning has implications for how we design schools for equity. In her foundation work, she has worked to bring a deep equity lens to grantmaking, and has spearheaded innovative funding opportunities rooted in the promise of research to support more equitable education systems. She is a member of the American Academy of Arts & Sciences and the National Academy of Education, and is a Fellow of the American Educational Research Association. She is a Past President of the American Educational Research Association and serves on the board of Sage Publications, the National Equity Project, and the UC Berkeley Board of Visitors.

Michelle Odemwingie is the chief executive officer at Achievement Network. Michelle joined ANet nearly a decade ago as a coach and has since held roles as chief of school and system services and chief of staff, among others. This includes spearheading ANet's Breakthrough Results Fund in partnership with five school districts across the country. Through her work at ANet and in her local community, Michelle maintains a deep personal commitment to educational equity and ensuring all students are able to learn and thrive. A recognized strategic advisor and policy advocate for the future of assessments, she plays a key role in shaping the national conversation around instructional improvement. Michelle actively engages in education policy and system-level transformation, advising districts, policymakers, and nonprofit leaders on instructional strategy, assessment innovation, and equitable access to high-quality materials. Prior to joining ANet, she spearheaded the ThinkMath team in California and DC, supporting instructional leaders around math enrichment and intervention programs, as well as supporting secondary math teachers through TNTP and Teach for America. Michelle began her career as an educator teaching math in the District of Columbia and is a graduate of Stanford University.

Maria Elena Oliveri is a Research Associate Professor of Engineering Education at Purdue University, working on the SCALE program. She is dedicated to developing innovative and equitable assessment approaches that prepare learners for professional practice. Her research focuses on improving assessment methods in engineering learning contexts, with particular attention to fairness, culturally and linguistically relevant assessment, assessing complex engineering competencies, and aligning assessments with evolving workforce needs. She has extensive expertise in the development of simulations, performance-based assessments, and the assessment of complex professional skills. She has played a leading role in shaping international assessment standards and best practices. She served as Chair for the International Test Commission's (ITC) Guidelines for the Fair and Valid Assessment of Linguistically Diverse Populations and as a steering committee member for the ITC Technology-Based Assessment Guidelines. She has authored various guidelines and standards in the field of assessment and has published over 100 peer-reviewed journal articles and conference papers. She is a multilingual researcher and speaks Spanish, French, and Italian. Her research continues to advance equity and effectiveness in education and workplace readiness.

Saskia Op den Bosch is co-founder of RevX, where she leads R&D strategy and spearheads the development of our innovative assessment system. She brings 14 years of experience as an educational researcher, strategist, and peer-reviewed author, creating environments that foster a strong sense of self and community, intellectual growth, and real-world impact. Previously, she led R&D for Getting Ready for School, integrating SEL into early literacy across NYC Head Start centers, and coached grantees at Character Lab on translating research into classroom practice. As Partner of R&D at Transcend, she built the R&D blueprint that secured large-scale federal funding for the Whole Child Model. Saskia holds a B.S. in Psychology from Carnegie Mellon and an M.A. in Quantitative Methods from Columbia. Committed to reimagining assessment as a catalyst for growth, she ensures learning environments evolve alongside young people—equipping learners to step into their purpose and create meaningful impact.

Dr. V. Elizabeth Owen is an expert in game-based learning analytics, with over 20 years experience in the learning sciences and education. At Age of Learning, she specializes in optimizing adaptive learning systems through applied AI and machine learning. Previously, she worked as a researcher and data scientist with Google, GlassLab Games at Electronic Arts, Inc. (EA) and LRNG by Collective Shift, after earning a Ph.D. in Digital Media (Learning Analytics focus) from the University of Wisconsin-Madison. Dr. Owen's doctoral work was based at the Games+Learning+Society (GLS) center, which launched collaborations with EA, Zynga, and PopCap Games using game-based Educational Data Mining. Dr. Owen spent a decade as a K–12 educator and was a founding teacher at the Los Angeles Academy of Arts and Enterprise charter school. She holds a BA from Claremont McKenna College.

Trevor Packer is the head of College Board's Advanced Placement Program. In rigorous classes that range from calculus to studio art, Advanced Placement provides high-quality coursework and the opportunity for college credit to more than 3 million students every year. With a deep love for literature, Trevor spent his time prior to the College Board working in academia. He has taught composition and literature at the City University of New York and Brigham Young University.

Roy Pea is David Jacks Professor of Education & Learning Sciences at Stanford University, Graduate School of Education, and Computer Science (Courtesy). His extensive publications in the learning sciences focus on advancing theories, research, tools and social practices of technology-enhanced learning of complex domains. He founded and directs Stanford's Ph.D. program in Learning Sciences and Technology Design. He is a Fellow of the American Academy of Arts and Sciences, National Academy of Education, Association for Psychological Science, the American Educational Research Association, and The International Society for the Learning Sciences. His most recent books include Learning Analytics in Education (2018), The Routledge Handbook of the Cultural Foundations of Learning (2020), and AI in Education: Designing the Future (2023). He is co-author of the National Academy of Sciences books: How People Learn (2000), and Planning for Two Transformations in Education and Learning Technology (2003). His most recent research involves studies of appropriate roles for Generative AI in augmenting writing and its development, computer science education, virtual reality storytelling, and culturally responsive science learning with augmented reality. In 2018 he received an Honorary Doctorate from The Open University. He won the McGraw Prize for Learning Sciences Research in 2022.

James W. Pellegrino is Emeritus Professor of Psychology and Learning Sciences and Founding co-director of the Learning Sciences Research Institute at the University of Illinois Chicago. His research and development interests focus on children and adults thinking and learning and the implications of cognitive research and theory for assessment and instructional practice. He has published over 350 books, chapters, and articles on cognition, instruction, and assessment. His education research has been funded by the National Science Foundation, the Institute of Education Sciences, and private foundations. As Chair or Co-Chair of several National Academy of Sciences study committees he co-edited major synthesis reports on teaching, learning, and assessment, including *Knowing What Students Know: The Science and Design of Educational Assessment*. He previously served on the Board on Testing and Assessment of the National Research Council and is a lifetime member of both the National Academy of Education and the American Academy of Arts and Sciences. His service includes the Technical Advisory Committees of several states and consortia, as well as those of the College Board, ETS, OECD, and the National Center on Education and the Economy. He currently serves on the NAEP Validity Studies Panel and ETS' Visiting Panel on Research.

Mario Piacentini is a Senior Analyst in the Programme for International Student Assessment (PISA) at the Organisation for Economic Cooperation and Development (OECD). An expert in measurement, Mario leads the work on the PISA innovative assessments and the broader PISA Research & Development Programme. He works with international experts to design assessments of 21st century competences. His projects aim to expand the metrics we use to define successful education systems. He is one of the authors of the Global Competence (PISA 2018) and Creative Thinking (PISA 2022) assessment frameworks, and he is currently leading the development of the PISA 2025 assessment of Learning in the Digital World and PISA 2029 assessment of Media and AI Literacy. He also coordinates the development of an open-source platform to support the use of technology-enhanced, formative assessments in the classroom. Before joining PISA, he worked for the Public Governance and the Statistics Directorates of the OECD, the University of Geneva, the World Bank and the Swiss Cooperation. He has authored several peer-reviewed articles and reports and was co-editor of the OECD publication on Innovating Assessments to Measure and Support Complex Skills. Mario holds a Ph.D. in economics from the University of Geneva.

Mya Poe is Professor of English at Northeastern University. Her research focuses on writing assessment and writing development with particular attention to justice and fairness. For more than 20 years she has advocated against assessment practices that are based on weak construct models and that result in unnecessary barriers for students. She has published five books, including Learning to Communicate in Science and Engineering: Case Studies from MIT (CCCC 2012 Advancement of Knowledge Award); Race and Writing Assessment (CCCC 2014 Outstanding Book of the Year); Writing Placement in Two-Year Colleges: The Pursuit of Equity in Postsecondary Education(CWPA 2022 Book of the Year); and Rethinking Multilingual Writers in Higher Education: An Institutional Case Study. In addition to teaching undergraduate courses on writing research methods and scientific writing, she also teaches graduate courses on writing assessment and the teaching of writing. Her teaching and service have been recognized with the Northeastern University Teaching Excellence Award and the MIT Infinite Mile Award for Continued Outstanding Service and Innovative Teaching. She has directed writing programs at MIT and Northeastern University and has worked extensively with faculty across the U.S. to improve the teaching of writing. She is co-editor of the international writing research journal Written Communication.

Ximena A. Portilla is a Senior Research Associate at MDRC where she serves as Content Lead for the Measures for Early Success Initiative, shaping a vision for the assessment content covered by tools coming out of the initiative and connecting assessment developers to supports to ensure content is aligned with developmental science. Portilla is a developmental scientist whose research over the last 20 years has focused on a range of topics in the preschool and kindergarten years, including home visiting, school readiness, and classroom supports for early educators.

Dr. Elizabeth J. K. H. Redman is a Research Scientist specializing in technology and assessment at the National Center for Research in Evaluation, Standards, and Student Testing (CRESST). Her primary research interests include STEM education, educational games, and assessment design. Her recent research focus has been on incorporating assessment capabilities into educational games, including SEL and STEM games. She has experience running observational classroom studies, RCTs and evaluations of educational games.

Jeremy D. Roberts is Senior Director of Learning Technology for PBS KIDS, where he works closely with award-winning series such as Curious George, Molly of Denali, Work it Out Wombats!, and Lyla in the Loop to deliver innovative, educational, multi-platform media experiences to kids aged 2-8. Roberts' work focuses on demonstrating and optimizing the impact produced by PBS KIDS media at scale. One of Roberts' core initiatives is the PBS KIDS Learning Analytics research program, which uses safe anonymous gameplay data, analytics, statistical modeling, research, and AB testing, to systematically discover game design principles that best balance reach, engagement, and learning effectiveness. Roberts' work helps PBS KIDS improve its overall impact by feeding relevant insights directly into the design, production, packaging, and distribution of PBS KIDS media. Over the decades, Roberts has cultivated a deep strategic understanding of technology, and the fast-evolving nature of the media, entertainment, and learning landscapes. A physicist by training, Roberts' passion for discovery and innovation has driven his extensive involvement with leading-edge technologies, and continues to define his work as an executive, leader, strategist, and systems engineer. To keep things interesting, Roberts plays trombone with D.C. soul, ska, and reggae band The Pietasters.

Dr. Mary-Celeste Schreuder is the Director of Literacy at the Achievement Network (ANet), where she leads ANet's national rollout of the Rapid Online Assessment of Reading (ROAR) in collaboration with Stanford University. With 20+ years in education, including roles as a secondary ELA teacher, professor of teacher education, and literacy strategist, Mary has built deep expertise in adolescent literacy, assessment strategy, and writing pedagogy. She designs tools, leads professional learning, and equips coaches and system leaders to support striving readers through research-based, equity-centered solutions. Her scholarship has been published in journals like the *Journal of Adolescent & Adult Literacy*, and she holds a Ph.D. in Literacy, Language, and Culture from Clemson University.

David Sherer is Director, Future of Assessment, at the Carnegie Foundation. In this role, he leads the Skills for the Future initiative, in collaboration with colleagues at ETS, to create a robust, scalable suite of assessment and analytic tools that captures the full range of skills required for students to succeed in K-12, post-secondary education and beyond. David coaches educational leaders in the use of evidence in the improvement process, the development of indicators and measures, and the assessment of organizational health. He holds a master's degree and a doctorate (Ed.D.) from the Harvard Graduate School of Education.

Stephen G. Sireci, Ph.D., is Distinguished Professor and Executive Director of the Center for Educational Assessment in the College of Education, University of Massachusetts Amherst. He earned his Ph.D. in psychometrics from Fordham University and his master and bachelor degrees in psychology from Loyola College Maryland. Before UMass, he was Senior Psychometrician at GED Testing Service, Psychometrician for the CPA Exam and Research Supervisor of Testing for the Newark NJ Board of Education. He is known for his research in validity and fairness of educational tests, and for innovations in test development. He currently serves/has served on several advisory boards including the National Board of Professional Teaching Standards, Duolingo English Test, and technical advisory committees for Florida, Maryland, New Hampshire, New York, Montana, Puerto Rico, and Texas. He is a Fellow of American Educational Research Association and of Division 5 of American Psychological Association, and a lifetime member of the National Academy of Education. He is a past President of International Test Commission, Northeastern Educational Research Association, and National Council on Measurement in Education. His UMass honors include School of Education's Outstanding Teacher Award, Conti Faculty Fellowship, Public Engagement Fellowship, Outstanding Accomplishments in Research and Creative Activity Award, and the Chancellor's Medal. He also received the Messick Memorial Lecture Award from Educational Testing Service/International Language Testing Association. He serves on several editorial boards including Applied Measurement in Education, Educational Assessment, Educational Measurement: Issues and Practice, Educational and Psychological Measurement, Practical Assessment Research and Evaluation, and Psicothema.

Dr. Erica Snow is the Senior Director of People Science and Analytics and Early Career Recruiting at Roblox. Previously, she was Director of Learning and Data Science at Imbellus, a game-based assessment startup acquired by Roblox. She also worked at SRI international as the Lead Learning Analytics Scientist before joining Imbellus. Dr. Snow has over a decade of experience evaluating the implementation and impact of a variety of educational technologies(i.e., ITSs, MOOCs, LMS, and blended learning courses) within K-12, postsecondary education, and workforce training. Her work has been presented both domestically and internationally to both scientific and non-scientific colleagues and has been published in over 70 peer-reviewed publications. She holds a Ph.D. and MA in Cognitive Science from Arizona State University and a BA in Psychology from Ball State University.

Rebecca A. Stone-Danahy has served as College Board's Director of AP Art and Design since 2020, where she has spearheaded initiatives to support course growth and advocacy ensuring access to inquiry-based art education through assessment practices. She also led the transformation of a physical to digital annual AP Art and Design exhibit, enhancing the visibility of diverse and high-quality student artworks. Rebecca's leadership in K-12 education spans roles from visual arts educator to fine arts administrator where she focused on inquiry-based visual art pedagogy, curriculum design, fine arts programming, and teacher mentorship. She is a strong proponent of integrating technology into education and was pivotal in launching one of the first online distance learning programs and museum collaborations between the North Carolina Virtual Public Schools and the North Carolina Museum of Art. Rebecca holds an MA in Art Education from Miami University in Oxford, Ohio, an M.Ed. in Secondary School Administration and an Ed.S. in Educational Leadership-School Superintendent from The Citadel in Charleston, SC, and an Ed.D. in Educational Systems Improvement Science from Clemson University in Clemson, SC. Rebecca's dissertation focus aimed to improve access and equity to inquiry-based visual art education for Title I school students in South Carolina.

Rebecca Sutherland, Ed.D., is the Associate Director of Research at Reading Reimagined, a funded program of the Advanced Education Research and Development Fund, where she leads a portfolio of research projects investigating the root causes of reading struggles among older students and instructional resources designed to address them. Rebecca has worked with K-12 public education data for over two decades to generate actionable knowledge for state and local agencies, and nonprofit organizations. She has taught ESL and reading in public schools in Japan and New York, and adult literacy in New York and Massachusetts. Rebecca holds a doctorate in Human Development and Psychology from the Harvard Graduate School of Education, a masters degree in Educational Psychology from the New York University Steinhardt School of Education, and a B.A. in history from Barnard College.

Natalya Tabony is Executive Director of AP Strategy and Analytics at the College Board. She leads a team focused on shaping program and product strategies that help more students access—and succeed in—Advanced Placement. Her work centers on using data and research to guide thoughtful decisions about how to strengthen the AP program and ensure it meets the needs of students and schools. Natalya began her career as a consultant with Parthenon-EY's education practice, where she worked on strategy and growth projects for school systems, universities, and philanthropic foundations. She later served as Director of Operations at a middle school in the Uncommon Schools network in Brooklyn, overseeing all aspects of daily operations. Across roles, she's been drawn to questions about how to improve schools and create more moments where students can discover what they're capable of. She emigrated from Russia to the U.S. as a child and grew up believing in the power of education to shape opportunity. Natalya holds a BA from Dartmouth College and an M.B.A. from the Kellogg School of Management. She lives in New York City with her husband and two young children.

Carrie Townley-Flores is the Director of Research and Partnerships for the Rapid Online Assessment of Reading (ROAR) at Stanford University. She holds a Ph.D. in Education Policy from Stanford. Her research focuses on reading assessment and related policies and practices that mitigate racial, ethnic, and economic inequality in the U.S. She joined the ROAR project with extensive experience working with schools, both in the classroom and in academic research-practice partnerships. Carrie taught English Language Arts at secondary schools in Michigan and New Hampshire and a primary school in Helsinki, Finland. She holds a B.A. in English and Education from University of Michigan.

Eric M. Tucker is the President and CEO of the Study Group, which exists to advance the best of artificial intelligence, assessment, and data practice, technology, and policy. He has served as President of Equity by Design, Superintendent and Executive Director of Brooklyn Laboratory Charter Schools, CEO of Friends of Brooklyn LAB, Cofounder of Educating All Learners Alliance, Executive Director of InnovateEDU, director at the Federal Reserve Bank of New York, and Cofounder and Chief Academic Officer of the National Association for Urban Debate Leagues. As an entrepreneurial, strategic, and impact-focused leader, Eric has over 25 years of experience building catalytic partnerships in education, securing over $300 million of investments for enterprises and initiatives that have transformed outcomes for learners and educators. Eric has expertise in measurement and assessment system innovation, participatory and advanced R&D, analytics, and human infrastructures for improvement and co-edited *The Sage Handbook of Measurement*. He earned a doctorate and a masters of science in measurement sciences from the University of Oxford and bachelors degrees from Brown University. Eric served as an ETS MacArthur Foundation Fellow with the Gordon Commission on the Future of Assessment in Education. He served as a Senior Research Scientist at the University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

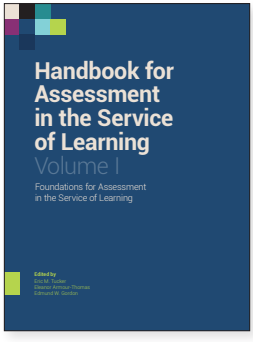
Alina A. von Davier is a researcher, innovator, and an executive leader with over 20 years of experience in EdTech and in the assessment industries. She is the Chief of Assessment at Duolingo, leading the Duolingo English Test research and development area. She is the Founder and CEO of EdAstra Tech. She is an American Educational Research Association (AERA) Fellow and serves as an Honorary Research Fellow at University of Oxford, and a Senior Research Fellow Carnegie Mellon University. Her research spans computational psychometrics, machine learning, and education. Dr. von Davier's work has been widely recognized in the academic community. She received the Brad Hanson award twice from National Council on Measurement in Education (NCME) for her pioneering work on computational psychometrics, and her work on adaptive testing. She received ATP's Career Award for her contributions to assessment. She was a finalist for the Innovator award from the EdTech Digest. The AERA awarded her the Division D Signification Contribution Educational Measurement and Research Methodology Award for her publications "Computerized Multistage Testing: Theory and Applications" (2014) and an edited volume on test equating, "Statistical Models for Test Equating, Scaling, and Linking" (2011).

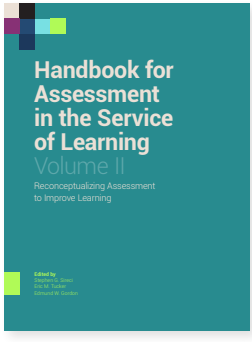
Kevin Yancey is a Senior Staff AI Researcher at Duolingo, leading the engineering and AI functions for Research & Development on the Duolingo English Test.As an expert software engineer and AI researcher who has also taught and studied abroad in two foreign countries, he is passionate about the applications of technology to second language learning and assessment. His work in AI specializes in the field of Natural Language Processing (NLP), where he has made innovative contributions to automatic readability estimation, automatic writing evaluation, and estimating item response theory (IRT) item parameters for L2 assessments using explanatory models with NLP features.

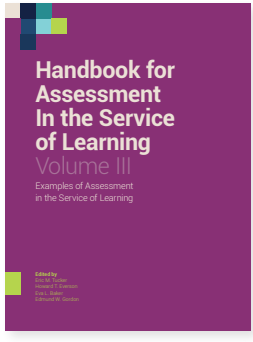
Jessica W. Younger, Ph.D., is an educational neuroscientist dedicated to developing effective interventions that empower learners to reach their full potential. With over a decade of experience, her work explores how individual differences shape learning, leveraging advanced statistical modeling and large-scale data analysis to personalize education. Currently, as Senior Manager of Research Products at PBS KIDS, Younger leads efforts to optimize educational content through innovative research tools, data-driven insights, and experimental platforms. Throughout her career, she has led multidisciplinary teams in designing research platforms, digital assessments, and large-scale studies that examine cognitive development and learning variability. Her work spans executive function, digital interventions, and personalized learning, with a focus on translating research into actionable insights for educators, technologists, and policymakers. By integrating neuroscience, data science, and education, Younger remains committed to advancing the understanding of how people learn best—ensuring that educational approaches are inclusive, evidence-based, and tailored to the needs of diverse learners.

Constance Yowell is senior advisor to the provost for special projects at Northeastern University. She previously served as senior vice chancellor for educational innovation, where she led the university's Center for Advancing Teaching and Learning Through Research, the University Honors Program, Undergraduate Research and Fellowships, Employer Engagement and Career Design, the Global Experience Office, Peer Tutoring, Self-Authored Integrated Learning, and the PreMed and PreHealth Advising Program. Before joining Northeastern, Yowell served as executive vice president of Southern New Hampshire University where she oversaw community engagement and outreach, with a focus on engineering a stackable, personalized learning approach for low-income, first-generation learners. Yowell began her career as an associate professor at the University of Illinois after serving as a policy analyst in the New York City school system and the U.S. Department of Education. Her research and policy work have focused on the deep disparities in local and federal education systems, particularly for African American and Latinx students, and she has written prolifically on the impact of educational policies and equity on student outcomes. Yowell holds a Ph.D. in child and adolescent development from Stanford University and a bachelor's degree from Yale University.

Handbook for Assessment in the Service of Learning Series







UMassAmherst

University Libraries

Volume I of the *Handbook for Assessment in the Service of Learning* offers a theoretical and research-grounded vision for transforming educational assessment into a catalyst for learning. Drawing on contemporary learning sciences, measurement theory, and improvement science, the volume is organized into three sections that offer principled design and conceptual frameworks for integrating assessment with teaching and learning; ground assessment in the social, cultural, and developmental nature of how people learn; and examine how emerging technologies like artificial intelligence might enrich balanced assessment practices while upholding technical requirements of validity: fairness, scientific soundness, utility and credibility. Informed by the vision of the Gordon Commission for the Future of Assessment in Education, this volume explores rethinking assessment as an integral component of pedagogy that informs the processes for learning and its improvement over time rather than just a final evaluation of the status of learning achieved. It provides the foundations for building assessment systems that are human-centered, just, and truly in the service of every learner.