Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 24, September 2022 ISSN 1531-7714

Developing and Refining a Model for Measuring Implementation Fidelity for an Instructionally Embedded Assessment System[1,2,3]

Jennifer L. Kobrin, *University of Kansas*
Meagan Karvonen, *University of Kansas*
Amy Clark, *University of Kansas*
W. Jake Thompson, *University of Kansas*

We developed a six-step iterative process for developing and evaluating a model of implementation fidelity appropriate for use in an instructionally embedded assessment system. Our work explicitly connects the literature on theories of actions for assessment systems with the implementation fidelity literature originating from the program evaluation field. The steps include (a) developing a logic model identifying critical and optional implementation components; (b) identifying process data and indicators from the assessment system to represent each component; (c) developing hypotheses about expected patterns in the indicators representing different levels of implementation fidelity and identifying criteria for defining implementation levels; (d) conducting analyses to test the hypotheses; (e) using the results to refine the indicators and criteria; and (f) evaluating strength of the evidence and identifying gaps. This process facilitates measuring action mechanisms and making and testing hypotheses about how critical implementation components are related to intended outcomes of an assessment. Studying implementation fidelity for assessment systems can help us better understand how teachers use assessment results and where additional support may be needed. This work can also help evaluate the extent to which instructionally embedded or formative assessments are implemented as intended and that all students are provided with sufficient opportunity to demonstrate what they have learned.

Keywords: formative assessment, instructionally embedded assessment, implementation fidelity, theory of action, logic model

Introduction

Instructionally embedded assessments are designed to help teachers understand students' learning by integrating ongoing data collection with classroom instruction (Pellegrino et al., 2016; Swinburne Romine & Santamaria, 2016). By design, instructionally embedded assessments do not merely serve as an

[1] Jennifer L. Kobrin https://orcid.org/0000-0002-5143-8884; Meagan Karvonen https://orcid.org/0000-0003-2071-2673; Amy K. Clark https://orcid.org/0000-0002-5804-8336; W. Jake Thompson https://orcid.org/0000-0001-7339-0300

[2] Statements and Declarations Funding: No funding was received for conducting this study. Conflicts of Interest/Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

[3] Acknowledgements. The authors would like to acknowledge Jeffrey Hoover for his data analysis contributions and Thanos Patelis for his review of an earlier draft of this manuscript.

indicator of student achievement; they are designed to lead directly to action on the part of the teacher and student. In cases where assessment systems are intended to serve as agents for action, it is incumbent upon the test developer to develop a theory of action documenting what needs to be in place for the desired effects to occur, as well as the ways in which improper implementation may lead to unintended negative consequences (National Council on Measurement in Education [NCME], 2018). A theory of action for an assessment system includes the assessment's intended effects, assessment components and their rationale, interpretive claims, action mechanisms, and potential unintended negative effects and what will be done to mitigate them (e.g., Bennett, 2010; Clark & Karvonen, 2020; Clark & Karvonen, 2021; Formative Assessment for Students and Teachers State Collaborative on Assessment and Student Standards [FAST SCASS], 2018; Gholson & Guzman-Orth, 2019; Wylie, 2017).

Action mechanisms, or the ways in which the claims are connected, are particularly important in a theory of action because these mechanisms connect an assessment system's components to the assessment's intended effects. In other words, the action mechanisms represent expectations and assumptions about the things that teachers and students must do beyond merely administering or taking an assessment to achieve the assessment's intended effects. For example, one of the action mechanisms and intended outcomes in the FAST SCASS (2018) theory of action is

> when teachers implement formative assessment in intentional and ongoing ways, and are more confident and satisfied, the implementation of quality teaching practices increases for both experienced and novice teachers (including preservice teachers). (p. 13)

In Bennett's (2010) theory of action, one of the action mechanisms is "teachers and students use . . . inferences [derived from formative assessment] to adjust instruction" (p. 72). Because action mechanisms are directly associated with an assessment's impact, it is important to measure the extent to which the action mechanisms take place (NCME, 2018).

The concept of implementation fidelity, common in evaluation research, can be used to guide the evaluation of action mechanisms in an assessment's theory of action. *Implementation fidelity* is "the extent to which an enacted program is consistent with the intended program model" (Century et al., 2010, p. 202). Measuring implementation fidelity has shown promise in bringing to light factors that may hinder program adherence, understanding how the quality of implementation impacts outcomes, and identifying the types of supports needed to ensure better implementation (Dhillon et al., 2015). Such studies can also document deviations from or variations within an intended model. Teachers may need to adapt a program or innovation to meet their students' needs and according to different instructional settings (Dusenbury et al., 2003).

Implementation fidelity has been defined in various ways, but most definitions reference a comparison between the critical components of a program's intended model and the components that are present when the program is actually enacted (e.g., Century et al., 2010). Researchers have used various approaches to identify critical components of a program and measure the extent to which these components are implemented. Dane and Schneider (1998) identified five dimensions of implementation fidelity to be measured: (a) adherence or the extent to which program components are delivered as designed; (b) exposure/dose, such as the number and length of sessions or the frequency with which program components are implemented; (c) quality of program delivery, which includes qualitative aspects such as enthusiasm and preparedness of the implementer; (d) participant responsiveness; and (e) program differentiation, or whether participants received only the planned interventions. Other approaches to measuring implementation fidelity include the critical components approach (e.g., Bond et al., 2000), the structure and process approach (Mowbray et al., 2003), and the use of fidelity frameworks such as the concerns-based adoption model and levels of use (Hall & Hord, 1987).

Century et al. (2010) combined existing approaches and developed a conceptual framework applicable across multiple programs and contexts. The framework includes two broad organizational categories, each with two subcategories of critical components: *structural*, which includes *procedural* and *educative* components; and *instructional*, which includes *pedagogical* and *student engagement* components. The structural components represent what a teacher needs to do (procedural) and know (educative) to administer

a program or intervention with fidelity, and the instructional components represent the actions, behaviors, and interactions teachers (pedagogical) and students are expected to engage in to implement a program or intervention with fidelity.

Although measuring implementation fidelity is common in educational and health evaluation, it is not prevalent in educational assessment. Grisham-Brown et al. (2008) and Reed and Sturges (2012) examined what they termed *assessment fidelity*, defined as the degree to which test administrators conform to established assessment procedures and protocols during administration. Fidelity to intended assessment administration procedures is analogous to Century et al.'s (2010) procedural component.

There are also a few studies examining teachers' implementation of formative assessment as part of ongoing instruction (Furtak et al., 2008; Hondrich et al., 2016; Mills & Ragan, 2000). Furtak et al. (2008) focused on two aspects of implementation fidelity that are based on Dane and Schneider's (1998) framework: adherence and quality of delivery (which correspond to Century et al.'s (2010) structural–procedural and instructional–pedagogical components, respectively). In a similar study, Hondrich et al. (2016) examined teachers' implementation of a curriculum-embedded science assessment that was based on students' short written tasks. Teachers were expected to provide individualized written feedback on the assessments and adapt instruction by assigning differentiated worksheets according to student assessment performance. The researchers identified three critical components of the formative assessment: assessment, which corresponds to Century et al.'s (2010) structural–procedural component; feedback, and adaptive instruction, which both correspond to Century et al.'s instructional–pedagogical component. Feedback was measured by whether teachers provided written, individualized feedback on students' assessments, and adaptive instruction was measured by whether teachers assigned differentiated worksheets to students depending on their assessment results. Mills and Ragan (2000) modeled their approach after the concerns-based adoption model (Hall & Hord, 2006) and identified 15 different implementation components with five variations of implementation fidelity for each component. The components covered all four of Century et al.'s (2010) categories of critical components.

Although prior research on assessment administration fidelity and implementation fidelity of formative assessment provide useful information to evaluate an assessment's theory of action and validity argument, they do not cover the full scope of fidelity for an instructionally embedded assessment model. In an instructionally embedded assessment model, fidelity includes uses of assessment results (i.e., action mechanisms) to guide instruction. Some of the critical components measured in prior studies are similar to the notion of action mechanisms. Specifically, Hondrich et al.'s (2016) measures of feedback and adaptive instruction captured teachers' actions resulting from their use of assessment information. However, prior approaches have not situated implementation fidelity into a theory of action that facilitates making and testing cause and effect (if/then) hypotheses about how critical components are interrelated and how they are intended to lead to assessment outcomes. Our work extends beyond current conceptions of assessment implementation fidelity to include action mechanisms.

The purpose of this paper is to illustrate a six-step process for developing and evaluating a model of implementation fidelity appropriate for use in an instructionally embedded assessment system. The process is grounded in the Dynamic Learning Maps (DLM) alternate assessments, a statewide assessment system for students with significant cognitive disabilities. We undertook an iterative approach to defining implementation fidelity criteria according to a logic model that draws from the assessment program's theory of action. We used the logic model to guide identification of indicators from the assessment system and conducted exploratory analyses to refine the indicators and criteria. The results from the exploratory analyses were used to test our theories and assumptions about intended use and inform additional data collection. As the field is considering increasingly flexible assessment systems, our fidelity indicators might be useful in other systems. Other assessment programs could also follow or adapt our six-step process to define their own indicators and gather evidence of the extent to which the assessment system was implemented as intended to lead to desired outcomes.

Methods

Context: Dynamic Learning Maps instructionally embedded assessment system

The purpose of the DLM alternate assessment is to measure alternate academic achievement standards in English language arts (ELA) and mathematics for students with the most significant cognitive disabilities who cannot demonstrate what they know on general education assessments, even with accommodations. States participating in the Dynamic Learning Maps (DLM) Consortium can choose between two assessment models, and five states use the instructionally embedded model. The instructionally embedded model has two 15-week administration windows, occurring during the fall and spring, respectively. After administration, the system updates to show student mastery of assessed levels for each tested standard. Summative reporting used for accountability purposes is based on all responses collected throughout the year.

States adopting the instructionally embedded model administer assessments on content standards of the teacher's choosing within blueprint constraints. The blueprints are organized by groups of conceptually related standards in each subject and grade. Teachers choose which standards to assess within the constraints. For example, the blueprint for third-grade ELA requires assessment on three of eight available standards related to determining the critical elements in a text. DLM assessments are short; they include between three and nine items measuring the standards at one of five levels that reflect varied complexity from the grade-level target, including three precursor skills and one successor skill. Teachers can choose the level of assessment for each standard. The system recommends a level based on a survey of the student's skills that is completed before assessment administration, but teachers may choose a different level.

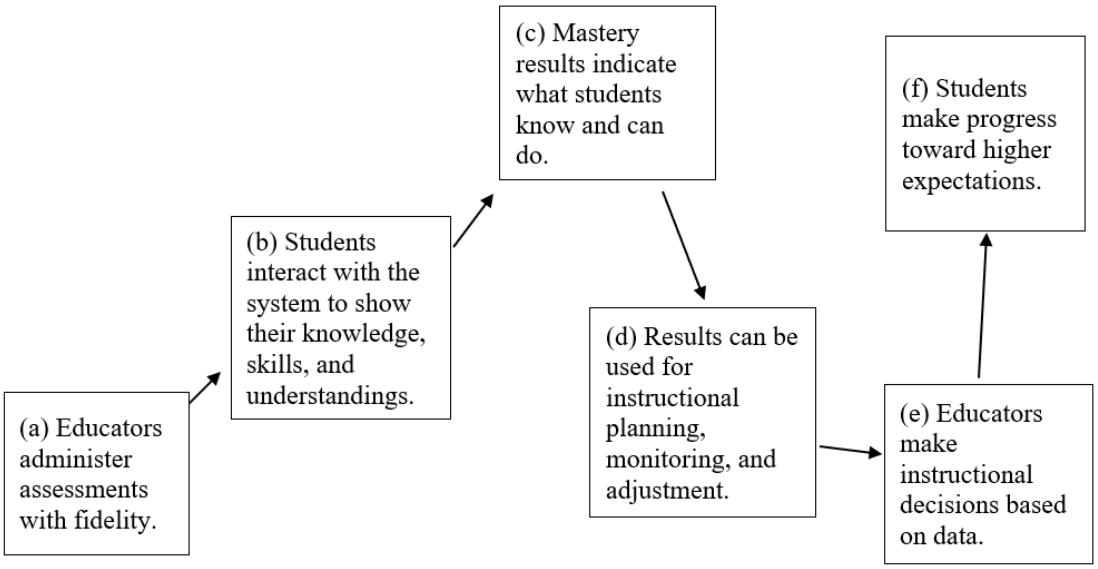
The teachers' choice of standards balances breadth of coverage with the flexibility to target individualized learning priorities. Choice of levels allows teachers to identify the complexity of the content for each standard that best matches the student's instructional level, so each student in this very heterogeneous population has meaningful access to the content and can meet the highest possible expectations. Thus, the assessments are teacher driven, allowing flexible selection of standards, complexity levels for assessment, and administration timing, with the expectation that the teacher covers blueprint requirements. Teachers also have the option to retest on the same standard or level and use the assessment system beyond the minimum requirements in the blueprint. Teachers receive annual training and have access to numerous resources to support their use of assessments as intended.

The DLM theory of action represents a causal model for how DLM assessments are intended to achieve desired long-term outcomes and explains how the desired change is expected to occur. One of the claims in the DLM assessment's theory of action (A in the theory of action excerpt in Figure 1) is that educators administer assessments with fidelity (Clark & Karvonen, 2021). This claim encompasses assessment fidelity, in other words, the degree to which test administrators conform to established assessment procedures and protocols during administration (Grisham-Brown et al., 2008); the implementation fidelity claim also informs subsequent claims in the theory of action (claims B–F in Figure 1). Although the focus of the current work is on implementation fidelity, the DLM theory of action includes several other claims (not shown in Figure 1) that are equally important for the assessment to achieve its intended purposes, such as those related to assessment design, accessibility, and several others (Clark & Karvonen, 2021). We collect evidence to evaluate these claims in other ways (e.g., Clark & Karvonen, 2020).

We implemented a six-step iterative process to develop and evaluate our implementation fidelity model:

1. Develop a logic model identifying critical components for implementation aligned to the Century et al. (2010) framework.
2. Identify process data and indicators from the assessment system to represent each critical (and optional) component.
3. Develop hypotheses about expected patterns in the indicators to represent different levels of implementation fidelity and use those hypotheses to identify criteria for defining implementation levels.
4. Conduct analyses to test the hypotheses.

Figure 1. Excerpt of Dynamic Learning Maps Theory of Action



5. Use the results to refine the indicators and criteria.
6. Evaluate strength of the evidence and identify gaps.

Step 1: Develop a logic model identifying critical components for implementation

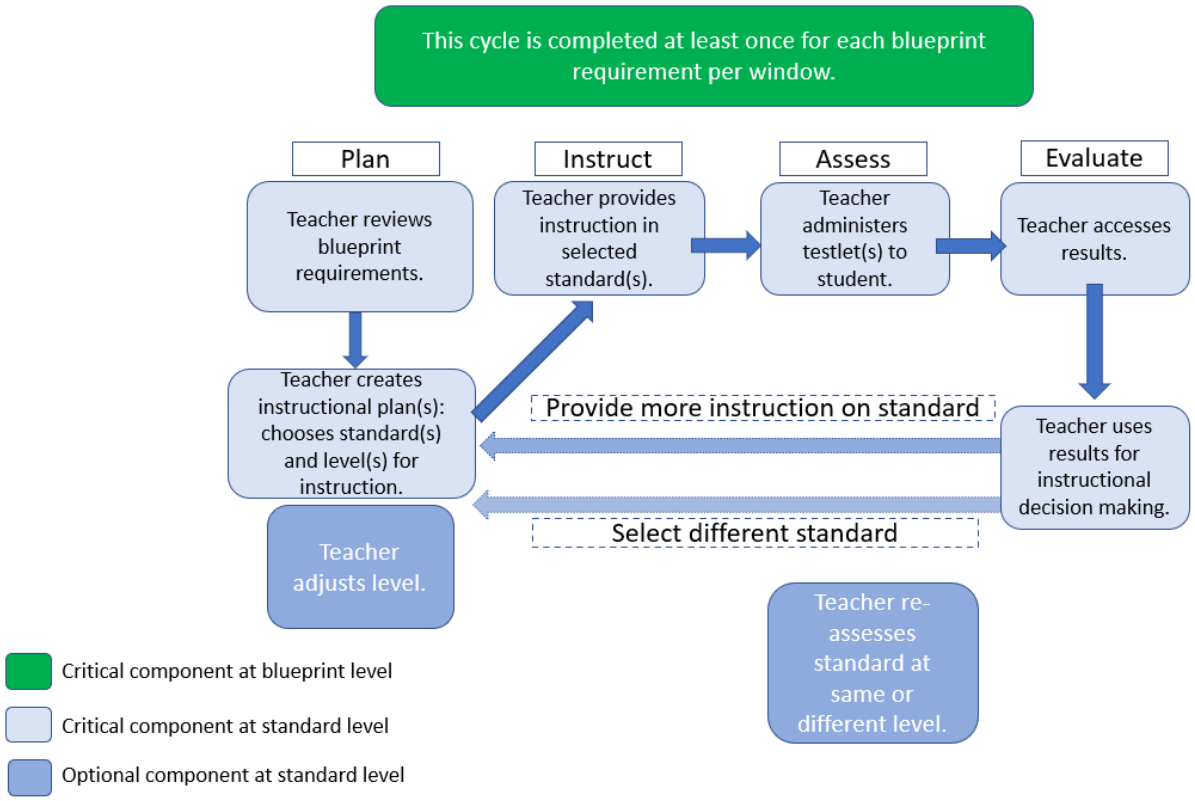
As a first step, we developed a logic model that explicitly defines the claim in the DLM assessment's theory of. action that. educators administer. the assessment with fidelity and articulates the structural and instructional components described by Century et al. (2010) that comprise implementation fidelity for the instructionally embedded assessments. Note that this logic model differs from the kind of logic model commonly used in program evaluation that typically specifies inputs, activities, outputs, and outcomes. This logic model (Figure 2) identifies the critical and optional components of implementation as defined by assessment manuals and trainings, other assessment documentation, and discussions with DLM staff; and was developed and refined by the authors in multiple rounds of review and discussion.

The DLM instructionally embedded assessments have critical components that must be in place to support implementation fidelity claims in the theory of action, as well as other optional components that offer teachers flexibility in supporting different instructional needs. Because the assessments allow flexibility in teacher choice at both the content-standard and overall-blueprint levels, the logic model reflects each of these levels. At the blueprint level, for each grade and subject, teachers are required to assess on a subset of available standards to meet blueprint requirements. If a student does not meet blueprint requirements, this indicates insufficient fidelity as the necessary breadth of content was not adequately assessed (i.e., insufficient construct representation). Thus, the blueprint level includes one critical component of blueprint coverage.

For each selected standard in the blueprint, teachers go through a cycle of instruction and assessment. The logic model at the standard level is based on the Test Administration Manual (DLM Consortium, 2019a), which describes five steps in the implementation cycle:

1. Select standard and level (Plan)
2. Provide instruction (Instruct)
3. Assess
4. Evaluate
5. Provide more instruction if needed, or select a different standard for instruction (Re-Assess)

Figure 2. Instructionally Embedded Assessment Logic Model



These five steps should be completed at least once for each standard the teacher selects during each of the two instructionally embedded assessment windows, and they may be repeated as needed at the discretion of the teacher.

The logic model identifies the components of instructionally embedded assessment implementation that should take place for each selected standard. The Plan, Instruct, and Assess steps are required components. In the Plan step, the teacher reviews the blueprint requirements and creates instructional plans in the online system by choosing the standard(s) and level(s) for instruction. However, for each standard, teachers may optionally adjust the system's recommended level if they believe a different level more appropriate for the student. In the Instruct step, teachers provide instruction on the selected standard(s). In the Assess step, teachers administer assessment(s) to the student.

The Evaluate and Re-Assess steps are currently optional. In the Evaluate step, the teacher views results in the online system and may use those results for instructional decision-making. The teacher may begin the cycle again for the same standard or a different standard by re-assessing the student at either the same or a different level. We recognize that these optional steps are key to the DLM theory of action because they represent teachers' use of assessment results to make instructional decisions and to act on those decisions (claims D and E in Figure 1). As we'll describe later, we are planning further research and development to fill this gap.

Table 1 shows the alignment of each component of the instructionally embedded logic model to Century et al.'s (2010) critical components. The logic model currently represents Century et al.'s structural–procedural and instructional–pedagogical components but does not represent the structural–educative and instructional–student-engagement components. The structural–educative component of the instructionally embedded assessment system is the required training that teachers complete before administering the

assessment. This is a separate claim in the theory of action (i.e., training strengthens educator knowledge and skills for assessing), which is evaluated with evidence related to the scope of training and requirements for passing the training, teacher-survey responses about their preparation to administer assessments, and other documentation. All teachers administering the DLM assessment must complete this training and demonstrate their knowledge on a posttest, so we assumed that all teachers met this critical component of implementation fidelity. The instructional–student-engagement component of the instructionally embedded assessment system is also a separate claim in the DLM theory of action (i.e., students interact with the system to show their knowledge, skills, and understandings). We currently collect evidence to evaluate this claim through test-administration observations, cognitive labs, assessment-completion rates, and analyses of teacher surveys and student-response patterns (DLM Consortium, 2019b; Karvonen et al., 2016).

Step 2: Identify process data and indicators from the assessment system to represent each component of the logic model

After developing the logic model, we identified process data and indicators representing the components of the logic model. We identified initial indicators based on descriptive analyses on participation and implementation using process data currently available from the assessment system representing teachers' decisions about which standards to assess and the timing and frequency of assessment (Clark et al., 2019).

For each stage of implementation in the logic model (Plan, Instruct, Assess, Evaluate, and Re-Assess), we identified indicators that provide evidence for the required and optional components. We examined the distributions and patterns of the indicators reported in Clark et al. (2019) and engaged in numerous discussions about which indicators were most important to evaluate claims in the theory of

Table 1. Implementation Fidelity Components for Dynamic Learning Maps (DLM) Instructionally Embedded Assessments

Step	Century et al. (2010) critical component	Required vs. optional	Description
Plan	Structural–procedural	Required	Completing blueprint requirements and creating instructional plans
	Instructional–pedagogical	Optional	Adjusting levels for assessment
Instruct	Instructional–pedagogical	Required	Providing instruction on selected standard(s)
Assess	Structural–procedural	Required	Administering assessment(s) according to published procedures
Evaluate	Instructional–pedagogical	Optional	Viewing reports and using results to make instructional decisions
Re-Assess	Structural–procedural	Optional	Administering assessment(s) according to published procedures
	Instructional–pedagogical	Optional	Choosing to re-assess students at the same level or a different level to assess mastery or progress
Outside system *	Structural–educative	Required	Completing required training to administer assessments
	Instructional–student engagement		

Note. * These critical components are separate claims in the DLM theory of action.

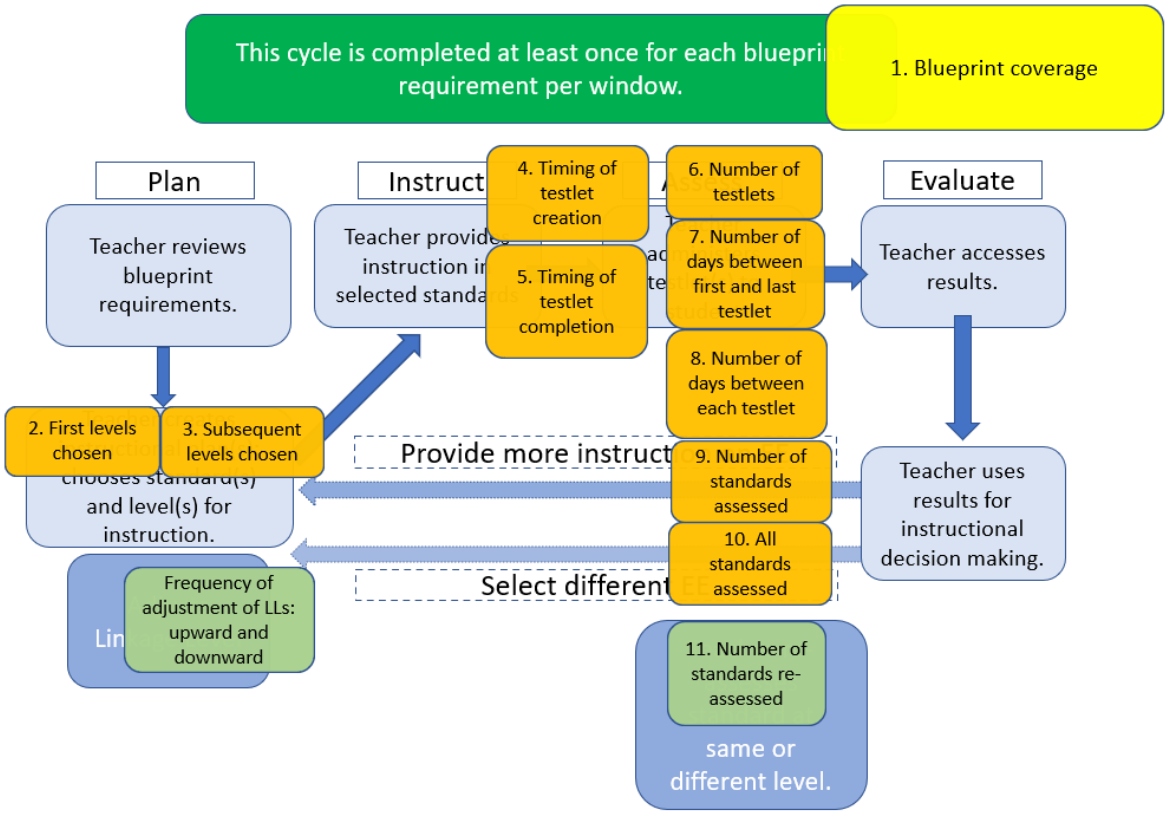
action, alternate hypotheses and what we could and could not infer from the data.

For example, we discussed the fact that many of the system indicators reflect structural–procedural fidelity, such as completing blueprint requirements, creating instructional plans, and administering the assessments according to published procedures. To support the theory of action, we knew that indicators on the instructional–pedagogical component were particularly important to represent teachers' use of the assessment for instructional decision-making. These indicators required us to make assumptions about patterns of data we might expect to see if teachers used the assessments as intended. For example, because we do not have a direct indicator of teachers' amount of instruction on a standard, we discussed the amount of lag time between assessments to expect if a teacher provided an adequate amount of instruction; we then confirmed our assumptions with colleagues who were former teachers of students with significant cognitive disabilities. We developed an algorithm (see Appendix A) to represent our hypotheses about the median number of days between assessment administrations that would suggest that teachers spent an adequate amount of time for instruction on a standard. Across grade levels and subjects, the lower boundaries ranged from 2.3 to 6.3 days, and the upper boundaries ranged from 7.3 to 12.6 days. We also discussed potential reasons teachers either do not meet blueprint coverage or assess on all possible standards, as well as the reasons teachers may adjust the levels of the assessments. As a result of these discussions, we settled on the list of indicators defined in Table 2 and shown overlaid on the logic model in Figure 3.

Table 2. Implementation Fidelity Indicators for Dynamic Learning Maps (DLM) Instructionally Embedded Assessments

Variable	Value	Description
Blueprint coverage	Not met Met Exceeded	For each subject and grade, teachers are required to test on a subset of all available standards to meet blueprint coverage requirements.
First levels chosen	Accepted recommended level Adjusted upward Adjusted downward	The DLM system recommends levels based on information entered about the student; however, teachers can adjust the level of the assessment for each standard either upward (more difficult) or downward (easier).
Subsequent levels chosen	Accepted recommended level Adjusted upward Adjusted downward	
Number of assessments	Range: 1–130	The system tracks the total number of assessments the student completes during the assessment window.
Number of days between first and last assessment	Range: 0–100.9	The number of days between the first and last assessment provides a rough estimate of the amount of time of instruction and assessment across all standards.
Median number of days between each assessment	Range: 0–62.9	The median number of days between each assessment indicates the spacing of assessment administration throughout the window.
All standards assessed	1 = yes; 0 = no	The system tracks whether the student assesses on all possible standards on the blueprint.
Number of standards re-assessed	Range: 0–26	Teachers have the option of re-assessing a student on one or more standards.

Figure 3. Indicators for Instructionally Embedded Assessment Implementation



Step 3: Develop hypotheses about expected patterns in the indicators to represent different levels of implementation fidelity and use those hypotheses to identify criteria for defining implementation levels

After the set of indicators were established in Step 2, we identified criteria for implementation fidelity (or a lack thereof) according to what we consider minimum requirements for intended use of the system and practices we believe support higher fidelity according to our theory of action. We used the Clark et al. (2019) findings as a source of initial evidence of implementation fidelity and to make inferences about the most prevalent patterns of system use. For example, the results suggested that most students meet or exceed blueprint coverage and that two prevalent of the window, which is not an intended practice, and students who spread testing throughout the window, with intervals between assessment administration of around five days (i.e., weekly administration).

We used our initial criteria to define three preliminary implementation levels (Levels 1, 2, and 3) to be used in subsequent analyses. In these early development iterations, we did not label the levels (e.g., sufficient, insufficient, strong) to avoid potential overinterpretation. The criteria for Level 3 were purposely stringent to identify cases that we believed truly exemplified exceptional use of the instructionally embedded assessments as they were designed to inform subsequent instructional decisions. Thus, we expected that Level 2 cases would represent teachers' meeting requirements but not taking full advantage of all system components to inform subsequent instruction. Level 1 is not meant to imply a lack of implementation fidelity but signals cases that we intend to investigate further.

Step 4: Conduct analyses to test the hypotheses and Step 5: Use the results to refine the indicators and criteria

Steps 4 and 5 were conducted in tandem in a few iterative cycles. We conducted a first round of analysis on the indicators, presented the findings to the assessment program's Technical Advisory Committee (TAC), and received feedback that prompted us to reevaluate our assumptions and revise the criteria. For example, one of the original criteria for Level 1 was

assignment and administration of all assessments either in the first or last 20% of the assessment window, suggesting a compressed schedule of assessment administration that did not allow adequate time for instruction. Our TAC suggested that we broaden the criterion to account for administration of all assessments within a 1-week period that can occur any time during the window.

Table 3 presents the refined criteria for Level 1 and Level 3 and rationales for their inclusion. Several indicators are part of the logic model but are not currently criteria to define levels of implementation. These indicators include the levels chosen for assessment and frequency of adjustments (upward and downward), total number of assessments, total number of standards assessed, and the timing of assessment completion across the assessment window. We had some hypotheses about how these indicators may vary by implementation level but do not yet have enough information on how these indicators relate to implementation fidelity or about the cutoffs we may use as thresholds for the implementation levels. For example, we hypothesized that students in Level 3 would assess more frequently and cover a larger breadth of standards than students in Levels 1 and 2. We did not have any firm hypotheses regarding the levels chosen for assessment or the frequency with which the teacher adjusts the level upward or downwards. On the one hand, excessive downward adjustment may introduce potential concerns about teachers providing opportunity for students to

Table 3. Instructionally Embedded Assessment Levels of Implementation: Criteria and Rationale

Level	Criterion	Rationale
1	Blueprint coverage not met.	The blueprint describes the minimum requirements for assessment.
	All assessments assigned and completed within a 1-week period.	Completing all assessments within a 1-week period may suggest the teacher did not provide adequate instruction on the standards that were assessed.
	Assessment of all possible standards.	Assessing all standards may represent a misunderstanding of requirements or not linking assessment and instruction.
3	Met or exceeded blueprint coverage.	The blueprint describes the minimum requirements for assessment. Teachers can choose to exceed the minimum requirements.
	Time between first and last assessment is at least 60 days.	The instructionally embedded window was 102 days in fall 2019; assessment over 60 days represents about 60% of the window, suggesting full use of the window for instruction and assessment on standards.
	Median days between assessments suggests adequate time for instruction.	After each standard is selected in the DLM system, we expect teachers to provide instruction on that standard so that students have maximum opportunity to demonstrate their knowledge, skills, and understandings on the assessment. If a student assesses on standards in close succession without a time gap, this may suggest that an adequate amount of instruction for each standard is not taking place.
	At least one standard is assessed more than once.	Re-assessment may indicate that teachers are reteaching material and providing students with additional opportunity to learn the content of the standard.

Note. Cases meeting any of the criteria for Level 1 were placed in that level. Cases must have met all criteria for Level 3 to be placed in that level. All cases not meeting the Level 1 or Level 3 criteria were placed in Level 2.

demonstrate their knowledge relative to the grade-level expectation or trying to game the system by selecting easier assessments to inflate student performance. However, there are legitimate reasons to adjust assessment levels to meet the needs of individual students. For instance, the system currently recommends one level for all standards, but students may be expected to have more-advanced skills on some standards than on others (e.g., algebra versus geometry).

To provide preliminary evidence to evaluate and refine the criteria, after the TAC meeting, we conducted a second round of analyses to examine differences in the indicators by implementation level. Data were obtained from the DLM instructionally embedded assessment system for the fall 2019 administration, representing 14,021 students in grades 3–11. A total of 13,995 students completed at least one assessment in ELA, and 13,704 completed at least one assessment in mathematics. The data represented 4,505 teachers, with an average of 3.1 students and a of two students per teacher (range of 1–24). As this research was largely exploratory, we generated descriptive statistics for the indicators in Table 1 for each implementation level and computed effect sizes of two students per teacher (range of 1–24). As this research was largely exploratory, we generated descriptive statistics for the indicators in Table 1 for each implementation level and computed effect sizes and odds ratios to examine differences among implementation levels. We used an effect-size calculator (Wilson, 2021) to compute odds ratios. We identified odds ratios equivalent to effect sizes of 0.20 or larger (Borenstein et al., 2011) to indicate where there were differences among implementation levels.

Using current criteria, 8,602 students (31.1%) were in Level 1, 18,945 (68.4%) were in Level 2, and 152 (0.5%) were in Level 3. In mathematics, a larger percentage of students were in Level 1 compared to ELA. Tables 4–6 show descriptive statistics for the indicators, as well as effect sizes and odds ratios for pairwise differences among the three implementation levels.

The results show that many of the variables differentiate the three levels according to our hypotheses. Most implementation indicators distinguish between Levels 1 and 3. The largest effect sizes are for the average days between the first and last assessment (2.59 and 2.96 in ELA and mathematics, respectively), average median days between each assessment (1.55 and 1.59 respectively), and average percentage of standards re-assessed (1.55 and 1.66 respectively). The days between first and last assessment and percentage of standards re-assessed were criteria for defining Level 3. Level 3 was characterized by a greater number of assessments, a longer testing window, greater spacing between assessments, and more frequent re-assessment of standards. In mathematics, teachers of Level 3 cases were also less likely than Level 1 cases to adjust levels upward and more likely to adjust downward. These same indicators differentiated Level 3 from Level 2, usually to a slightly lesser degree, with the exception of the number of standards re-assessed which showed greater differentiation between Levels 3 and 2.

Level 1 cases were more likely than other cases to complete all testing in either the first or last 20% of the assessment window, which may suggest a focus on completing assessment requirements rather than integrating assessment with the full semester of instruction. Alternately, it may suggest teachers are waiting until the end of the assessment window to assess to maximize instructional time and to provide instruction that connects across standards.

The effect sizes for the differences between Levels 1 and 2 show less differentiation between these levels, which may suggest a need to further refine the criteria; the largest difference is in the average days between first and last administered assessment ($d = 0.64$ in ELA and $d = 0.69$ in mathematics). In ELA, Level 2 cases were more likely than Level 1 cases to meet the threshold for the median number of days between assessments, suggesting adequate time for instruction. However, Level 1 and Level 2 cases had a similar percentage of standards re-assessed.

A key finding is that cases may not clearly lie in one implementation level; rather, teachers seem to exhibit a combination of practices, some that demonstrate higher fidelity to intended practice and others that do not. For example, 6.9% of the Level 1 cases had median days between assessments that suggests that the teacher spent adequate time for instruction on each standard, and 4.2% were re-assessed on at least one standard. This finding warrants further investigation and may influence subsequent development of our model.

Table 4. Descriptive Statistics for Implementation Indicators by Level and Subject

Implementation indicators		Level 1		Level 2		Level 3	
		ELA (*n*=3,570)	**Math (*n*=5,032)**	**ELA (*n*=10,329)**	**Math (*n*=8,616)**	**ELA (*n*=96)**	**Math (*n*=56)**
Blueprint coverage	% not met[a]	41.8	62.9	N/A	N/A	N/A	N/A
	% met[b]	50.7	27.1	79.9	72.0	60.4	50.0
	% exceeded[b]	7.5	10.0	20.1	28.0	39.6	50.0
First levels chosen	% accepted recommended level (*M*)	43.2	51.3	46.4	57.7	49.8	56.6
	SD	29.0	42.1	25.1	41.3	24.1	38.4
	% adjusted upward (*M*)	29.8	23.5	24.7	17.9	22.7	9.6
	SD	37.9	37.9	33.7	33.6	31.9	22.2
	% adjusted downward (*M*)	27.0	25.2	28.9	24.4	27.5	33.8
	SD	29.8	37.9	28.3	37.0	24.7	37.1
Subsequent levels chosen for same standard	% accepted recommended level (*M*)	31.7	31	31.8	30.4	35.4	35.6
	SD	33.2	39.2	34.0	38.0	38.7	43.8
	% adjusted upward (*M*)	41.6	44.6	37.6	43.1	34.2	27.2
	SD	39.6	43.5	39.1	43.1	39.9	40.9
	% adjusted downward (*M*)	26.7	24.4	30.7	26.5	30.4	37.2
	SD	35.0	37.7	36.6	37.7	40.7	44.2
Timing of assessment creation	% in first 20% of the window	4.0	5.0	9.0	7.0	4.0	0
	% in last 20% of the window	26.0	24.0	4.0	5.0	0	0
Timing of assessment completion	% in first 20% of the window	4.0	3.0	1.0	1.0	0	0
	% in last 20% of the window	28.0	30.0	13.0	16.0	0	0
	% all assessments completed within one week[a]	66.6	49.6	N/A	N/A	N/A	N/A
	% all assessments completed within two weeks	70.8	57.4	14.1	13.3	0	0
	% students whose median days between assessments suggests adequate time for instruction[b]	6.9	5.5	20.7	13.1	100	100
Number of assessments	*M*	6.8	7.2	7.4	7.9	9.6	9.7
	SD	2.9	4.3	3.3	4.4	2.1	1.6
Number of days between first and last assessment[b]	*M*	14.0	13.7	28.4	29.2	72.2	74.9
	SD	22.7	20.8	22.4	23.3	7.7	8.3
Median number of days between each assessment[b]	*M*	1.1	0.9	1.9	1.4	5.3	6.9
	SD	4.1	3.8	3.1	2.8	1.6	1.8
% all standards assessed[a]	*M*	0	2.9	N/A	N/A	N/A	N/A
	SD	N/A	16.7	N/A	N/A	N/A	N/A
% standards re-assessed[b]	*M*	4.2	3.4	3.9	3.3	28.5	27.1
	SD	15.6	14.2	14.3	13.9	19.5	19.9

Note. ELA = English language arts.

[a] Indicator used to define Level 1; [b] Indicator used to define Level 3.

Table 5. Effect Sizes for Continuous Implementation Indicators

Implementation indicators		Level 3 vs. Level 1		Level 3 vs. Level 2		Level 2 vs. Level 1	
		ELA	**Math**	**ELA**	**Math**	**ELA**	**Math**
First levels chosen	% accepted recommended level	0.23*	0.13	0.14	-0.03	0.12	0.15
	% adjusted upward	-0.19	-0.37*	-0.06	-0.25*	-0.15	-0.16
	% adjusted downward	0.02	0.23*	-0.05	0.25*	0.07	-0.02
Subsequent levels chosen	% accepted recommended level	0.11	0.12	0.11	0.14	0.00	-0.02
	% adjusted upward	-0.19	-0.40*	-0.09	-0.37*	-0.10	-0.03
	% adjusted downward	0.11	0.34*	-0.01	0.28*	0.11	0.06
Number of assessments		0.97*	0.58*	0.67*	0.41*	0.19	0.16
Days between first and last assessment [a]		2.59*	2.96*	1.96*	1.97*	0.64*	0.69*
Median days between each assessment [a]		1.55*	1.59*	1.10*	1.97*	0.24*	0.16
% standards re-assessed [a]		1.55*	1.66*	1.71*	1.71*	-0.02	-0.01

Note. ELA = English language arts.
[a] Indicator used to define Level 3. *Effect sizes of 0.20 or larger.

Table 6
Effect Sizes for Binary Implementation Indicators

Implementation indicators		Level 3 vs. Level 1		Level 3 vs. Level 2		Level 2 vs. Level 1	
		ELA[a]	Math	ELA	Math	ELA	Math
Timing of assessment creation	All in first 20% of the window	1.20*	---[d]	0.44*	---	2.70*	1.48*
	All in last 20% of the window	---	---	---	---	0.13	0.16
Timing of assessment completion	All in first 20% of the window	---	---	---	---	0.34*	0.34*
	All in last 20% of the window	---	---	---	---	0.39*	0.44*
	All assessments completed within one week[b]	---	---	---	---	---	---
	All assessments completed within two weeks	---	---	---	---	0.07	0.11
	Median days between assessments suggests adequate time for instruction[c]	---	---	---	---	3.53*	2.62*

Note. [a]English language arts; [b]Indicator used to define Level 1; [c]Indicator used to define Level 3. [d]Odds ratio could not be calculated because the frequency for one group = 0. *Odd ratios equivalent to effect sizes of 0.20 or larger.

Step 6: Evaluate strength of evidence and identify gaps

The last step of the process is to evaluate the strength of the implementation-fidelity evidence and identify gaps in data collection. This step may begin earlier in the process, when indicators are first identified. We do not currently use data on all teacher actions in the instructionally embedded assessment system. While we use data documenting when an assessment is assigned, we do not currently include data for when the instructional plan was first created or the amount of instructional time actually provided before administering an assessment. We do not currently track when teachers make changes to the instructional plan to adjust the standard or level for assessment before an assessment is administered; this adjustment may indicate an instructional decision based on a student's observed instructional level or may just reflect a mistaken assignment. We also do not currently include indicators on the extent and ways in which teachers access and use assessment results or on the instructional–student engagement critical components, which represent the actions and behaviors students are expected to engage in when participating in the assessment (Century et al., 2010).

Future steps

Because this study is based only on data from one fall assessment window, in the next stage of research we will replicate the analyses on a full year of data to cross-validate the findings. We will also more thoroughly explore alternative hypotheses and examine which indicators best differentiate between implementation levels. The indicators that continue to be nonsignificant in differentiating levels may be further refined or removed from the model. As our initial findings suggest that cases may not fit clearly into one implementation level, we will also explore developing profiles or types of instructionally embedded assessment use representing different combinations of critical components enacted to various degrees (Century et al., 2010), similar to Hall and Hord's (1987) innovation configurations. We will examine how different implementation patterns are associated with student outcomes to explore the decisions teachers make for different types of students and which implementation patterns have the greatest impact on student assessment results. Finally, we will collect qualitative data to further examine teachers' assumptions, motivations, and rationales for making various choices in the assessment system. We plan to conduct focus groups or surveys to better understand teachers' decisions and factors affecting their implementation and to evaluate the extent to which our inferences about use of the system aligns with practice. We will continue to evaluate and refine our indicators and criteria as we learn more from future research.

After the implementation-fidelity model is refined, we will develop plans for testing the action mechanisms in the theory of action. The implementation-fidelity claim feeds into a claim that "students interact with the system to show their knowledge, skills, and understandings" (claim B in Figure 1). Research to evaluate this action mechanism may include correlating implementation data to data collected from test-administration observations, as well as assessment-completion rates and aberrant-response analyses. We expect that cases with higher levels of implementation fidelity will have no concerns flagged in test-administration observations, will show better assessment-completion rates, and will be less likely to show aberrant-response patterns. Our ultimate goal in this research is to be able to routinely examine implementation fidelity and refine training and documentation as needed so teachers better understand how the assessments are intended to be used. Because states chose the instructionally embedded model to meet summative requirements and be instructionally useful, fidelity data also will be useful if future shifts in the model become necessary for philosophical or policy reasons.

Discussion

This study presents a six-step iterative approach to develop and evaluate a model of implementation fidelity for an instructionally embedded assessment system aligned with a theory of action. Other assessment programs could follow or adapt our six-step process to define their own indicators and gather evidence of the extent to which the assessment system was implemented as intended to lead to desired outcomes.

Our work explicitly connects the literature on theories of actions for assessment systems with the implementation fidelity literature originating from the

program evaluation field. Incorporating implementation fidelity frameworks into a theory of action facilitates measuring action mechanisms and making and testing if/then hypotheses about how critical implementation components are related to intended outcomes of an assessment. This approach is designed to support the selection and development of indicators that are aligned to intended uses of assessments to lead to desired changes in teacher and student behaviors and outcomes. Our approach is consistent with Cizek's (2020) recommendation to implement a research agenda that includes identification of theoretically related factors and empirical research to test causal claims in a theory of action.

Century et al.'s (2010) implementation fidelity framework guided the identification of indicators that are currently available in our assessment system and helped us evaluate where there are gaps. The indicators evaluated in this research study align most directly with Century et al.'s structural–procedural components; that is, they reflect assessment fidelity, including the basic steps teachers follow to set up instructional plans and administer the assessments. Some of the indicators address instructional–pedagogical components reflecting teacher actions and behaviors related to the instruction and assessment cycle that address the assessment system's theory of action. These instructional–pedagogical components are critical in embedded through-course and formative assessment systems as they represent teachers' use of assessment results for instructional decision-making.

Because these components are not directly measured in the DLM assessment system, we use indirect indicators to make inferences that need to be validated. For example, we used indicators on the amount of time between assessments to infer the amount of instruction on standards. We recognize that there are potential alternate hypotheses explaining teachers' decisions during their use of the assessment system. For instance, the amount of instruction needed on a standard is highly variable and dependent upon individual student needs and the level chosen for assessment (i.e., more instructional time may be needed for more difficult standards and higher levels). Additionally, there is wide heterogeneity in the DLM student population and legitimate reasons teachers may make different assessment and instructional choices for particular students. Teachers may also provide instruction on two or more standards simultaneously. Researchers developing hypotheses about instruction in programs where instruction is less individualized and variable across students should seek expert input on the degree of variability expected across classrooms.

Other assessment programs wishing to create a model of implementation fidelity can adapt our model with indicators that are appropriate for their program and its theory of action. Most of the indicators in our model are specific to the DLM and would not be appropriate for other programs. For example, since the overall number of required standards in the DLM varies by grade level and subject, we chose to include indicators on whether the student met blueprint requirements and was assessed on every possible standard rather than the percentage of standards assessed. In addition, because some of our indicators have skewed distributions we decided to use medians rather than means.

As the field considers implementation of increasingly flexible assessment systems that are designed to elicit change (e.g., Hedger, 2020), it will be important to develop approaches to systematically study variations and adaptations in implementation and their relationship to assessment outcomes. Studying implementation fidelity for assessment systems can help us better understand how teachers use assessment results and where additional support may be needed. This work can also help evaluate the extent to which instructionally embedded or formative assessments are implemented as intended and that all students are provided with sufficient opportunity to demonstrate what they have learned.

References

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement*, *8*(2–3), 70–91.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of*

Evaluation, *31*(2), 199–218. https://doi.org/10.1177/1098214010366173

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Taylor & Francis Group.

Clark, A. K., & Karvonen, M. (2020). Constructing and evaluating a validation argument for a next-generation alternate assessment. *Educational Assessment*, *25*(1), 47–64. https://doi.org/10.1080/10627197.2019.1702463

Clark, A. K., & Karvonen, M. (2021). Instructionally embedded assessment: A theory of action for an innovative system. *Frontiers in Education*. https://doi.org/10.3389/feduc.2021.724938

Clark, A. K., Thompson, W. J., & Karvonen, M. (2019). *Instructionally embedded assessment: Patterns of use and outcomes* (Technical Report No. 19-01). University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://dynamiclearningmaps.org/sites/default/files/documents/publication/IE_Usage_Report_2018.pdf

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical Psychology Review*, *18*(1), 23–45. https://doi.org/10.1016/s0272-7358(97)00043-3

Dhillon, S., Darrow, C., & Meyers, C. V. (2015). Introduction to implementation fidelity. In C. V. Meyers & W. C. Brandt (Eds.) *Implementation fidelity in education research* (pp. 22–36). Routledge.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research*, *18*(2), 237–256. https://doi.org/10.1093/her/18.2.237

Dynamic Learning Maps Consortium. (2019a). *Dynamic Learning Maps test administration manual 2019–2020*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.

Dynamic Learning Maps Consortium. (2019b). *2018-2019 technical manual update—integrated model*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://dynamiclearningmaps.org/sites/default/files/documents/publication/2018-2019_IM_Technical_Manual_Update.pdf

The Formative Assessment for Students and Teachers (FAST)/State Collaborative on Assessment and Student Standards. (2018). *An integrated approach to defining a system-level theory of action for formative assessment*. Council of Chief State School Officers. https://www.ccsso.org/sites/default/files/2018-06/Theory%20of%20Action%20for%20Formative%20Assessment%20Final.pdf.

Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P. R., Shavelson, R. J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education*, *21*(4), 360–389. https://psycnet.apa.org/doi/10.1080/08957340802347852

Gholson, M. L., & Guzman-Orth, D. (2019). *Developing an alternate English language proficiency assessment system: A theory of action* (ETS Research Report #ETS RR-19-25). ETS. https://doi.org/10.1002/ets2.12262

Grisham-Brown, J., Hallam, R.A., & Pretti-Frontczak, K. (2008). Preparing Head Start personnel to use a curriculum-based assessment. *Journal of Early Intervention*, *30*(4), 271–281. https://doi.org/10.1177/1053815108320689

Hall, G. E., & Hord, S. M. (1987). *Change in schools: Facilitating the process*. State University of New York Press.

Hall, G. E., & Hord, S. M. (2006). *Implementing change: Patterns, principles and potholes* (2nd ed.). Allyn and Bacon.

Hedger, J. (2020). States experiment with assessment through innovative pilots. *The State Education Standard*, *20*(3), 40–42. https://nasbe.nyc3.digitaloceanspaces.com/2020/09/Hedger_September-2020-Standard.pdf

Hondrich, A. L., Hertel, S., Adl-Amini, K., & Klieme, E. (2016). Implementing curriculum-embedded formative assessment in primary school science classrooms. *Assessment in Education: Principles, Policy & Practice*, *23*(3), 353–376. https://doi.org/10.1080/0969594X.2015.1049113

Karvonen, M., Swinburne Romine, R., & Clark, A. (2016, April 7–11). *Validity evidence to support alternate assessment score uses: Fidelity and response processes* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Washington DC, United States.

Mills, S. C., & Ragan, T. J. (2000). A tool for analyzing implementation fidelity of an integrated learning system. *Educational Technology Research and Development*, *48*(4), 21–41. https://doi.org/10.1007/BF02300498

National Council on Measurement in Education. (2018). *Position statement on theories of action for testing programs*. https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/Documents/NCME_Position_Paper_on_Theories_of_Action_-_Final_July__2018.pdf

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

Reed, D. K., & Sturges, K. M. (2012). An examination of assessment fidelity in the administration and interpretation of reading tests. *Remedial and Special Education*, *34*(5), 259–268. https://doi.org/10.1177/0741932512464580

Swinburne Romine, R., & Santamaria, L. (2016). *Instructionally embedded assessments* (Project Brief #16-01). University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Brief_16-01.pdf

Wilson, D. B. (2021). *Practical Meta-Analysis Effect Size Calculator*. https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php.

Wylie, E. C. (2017). *Winsight Assessment System: Preliminary theory of action* (ETS Research Report ETS RR-17-26). ETS. https://doi.org/10.1002/ets2.12155

Citation:
Kobrin, J. L., Karvonen, M., Clark, A., Thompson, W. J. (2022). Developing and refining a model for measuring implementation fidelity for an instructionally embedded assessment system. *Practical Assessment, Research, & Evaluation*, 27(24). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/24/

Corresponding Author:
Jennifer L. Kobrin
University of Kansas
Lawrence, Kansas, USA

Email: jennifer.kobrin [at] ku.edu

Appendix A.

Algorithm for Median Number of Days Between Assessment Administrations Suggesting an Adequate Amount of Time for Instruction

In order to set criteria for the median number of days between testlets that suggested adequate time for instruction on each standard, we calculated lower and upper boundaries based on the required number of testlets on the blueprint, which varies by grade level and subject. The lower boundaries were calculated as: [(*Number of testlets required for blueprint coverage) + 1]* to represent at least one standard assessed more than once, which is another criterion for Level 3 implementation. The upper boundaries were calculated as: *[(Number of standards on blueprint - 1)] *2* to represent re-assessment on all standards but the student was not assessed on every possible standard, since this is a criterion for Level 1 implementation. Once the lower and upper boundaries were determined for each blueprint, we calculated the range in the number of days between testlets as: *[{length of testing window} – {2 weeks} / {upper boundary}] to [{length of testing window} – {2 weeks} / {lower boundary}]*. We subtracted two weeks from the length of the testing window with the assumption that Level 3 implementation would not include assessment in the first two weeks of the assessment window to allow adequate time for instruction on the first standard selected for assessment. Across grade levels and subjects, the lower boundaries ranged from 2.3 to 6.3 days and the upper boundaries ranged from 7.3 to 12.6 days.