Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 28 Number 4, March 2023 ISSN 1531-7714

Applying a Contrasting Groups Standard Setting Methodology to a Large-Scale Performance Assessment Program Used for Accountability

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

Large-scale performance assessment programs are a longstanding reform tool. However, standard setting can be a challenge for assessment programs that use primarily non-standardized assessments. The purpose of this paper is to extend this field of research by explaining the standard setting methodology applied to one more recent instantiation of a state performance assessment program. The second purpose of this paper is to discuss the data quality control and quality assurance challenges experienced after five years of applying the standard setting method. Recognizing the burgeoning interest again in large-scale performance assessment programs, the goal and intended contribution of this paper is to inform future decisions about selecting appropriate standard setting methods and dealing with unanticipated challenges that may arise during implementation based upon the lessons learned from one program. It is likely that other large-scale performance assessment programs may face similar operational challenges, especially those that do not rely on standardized tests or standardized administration procedures to produce annual determinations of student proficiency or other scores used for accountability purposes. Assessment system designers can use the insights in this paper to consider standard setting methods and how those methods may need to be adapted to promote technical quality.

Keywords: Standard setting; Performance based assessment; Academic achievement; Accountability; Elementary secondary education

Introduction

Kentucky, Vermont, Maryland, amongst other states and districts experimented with large-scale performance assessment programs back in the 1980s and 1990s (Pecheone, Kahl, Hamma, & Jaquith, 2010; Stecher, 2010; Tung & Stazesky, 2010). More recently, the perceived lack of value state test results provide to educators and desire for assessment to drive higher quality teaching and learning in schools (amongst other reasons) has led to a resurgence of interest in the use of state-level performance assessments, as well as other types of assessments (Aurora Institute et al., 2021; Darling-Hammond & Adamson, 2014; Parsi & Darling-Hammond, 2015). These large-scale performance assessment programs have the potential to be used in multiple ways, including for school accountability purposes (Darling-Hammond & Snyder, 2015; Marion & Leather, 2015; Massachusetts Department of Education, 2020; New Hampshire Department of Education, 2016b) and for college admissions (Fine & Pryiomka, 2020; Guha, Wagner, Darling-Hammond, Taylor, & Curtis, 2018).

Yet there are many challenges in using results from performance assessments for a high-stakes use (Davey et al., 2015; Dunbar, Koretz, & Hoover, 1991; Koretz, Stecher, Klein, & McCaffrey, 1994; Ruiz-Primo, Baxter, & Shavelson, 1993; Shavelson, Baxter, & Gao, 1993; Tung & Stazesky, 2010; Yen & Ferrara, 1997). High-stakes use means that there are notable

consequences for some group based on the results. Standardized accountability tests are often considered high stakes for districts and schools if their results influence school ratings, educational funding, accreditation, property values in a community, etc. For students, performance on a high-stakes assessment may impact consequential decisions such as graduation, admission to college, scholarship awards, promotion, etc.

At the top of the list of challenges for large-scale performance assessment programs intended to be used in high-stakes ways is issues related to technical quality—specifically, concerns around the reliability and validity of scores for particular uses resulting from student responses, products, or performances (Pecheone et al., 2010; Tung & Stazesky, 2010). One aspect involved in validation processes related to score interpretations and uses for any educational assessment is using appropriate standard setting methodologies to designate cut scores that define levels of performance (AERA, APA, & NCME, 2014; Cizek, 2011; Kane, 2013). Previous large-scale performance assessment reforms created standard setting methods that supported non-standardized assessment information. For example, the ID Matching method was created for the Maryland School Performance Program (Ferrara & Lewis, 2011), and the Body of Work method arose out of the work with portfolios and other collections of evidence with complex performance tasks in Kentucky and other clients of Advanced Systems (later Measured Progress/Cognia) (Kingston & Tiemann, 2011).

The purpose of this paper is two-fold. First, to extend this field of research by explaining the standard setting methodology applied to one more recent instantiation of a state performance assessment system that relies exclusively on local classroom assessment information (including performance assessments) and teacher judgments to determine student proficiency for accountability purposes. The second purpose of this paper is to discuss the data quality control and quality assurance challenges experienced after five years of applying the standard setting method. Recognizing the burgeoning interest again in large-scale performance assessment programs, the goal and intended contribution of this paper is to inform future decisions about selecting appropriate standard setting methods and dealing with unanticipated challenges that may arise during implementation based upon the lessons learned from one program. It is likely that other large-scale performance assessment programs may face similar operational challenges related to standard setting, especially those that do not rely on standardized tests or standardized administration procedures to produce annual determinations of student proficiency or other scores used for accountability purposes.

This paper is organized as follows. The background section provides a detailed explanation of the state assessment system described in this paper in order to contextualize the choice for the contrasting groups standard setting methodology and what information is available to use within standard setting. The next section explains the key factors influencing standard setting in this state context and explains how the contrasting groups standard setting methodology was applied. Challenges related to applying the standard setting methodology are then addressed in detail related to data quality control and data quality assurance with embedded explanations of solutions applied and lessons learned. The paper then concludes with the significance of this paper for expanding understanding of standard setting methods and challenges related to state assessment systems that rely on non-standardized information.

Background

The state assessment system described in this paper—New Hampshire's Performance Assessment of Competency Education (PACE)— was a federally approved state assessment system under Section 1204 of the *Every Student Succeeds Act* (ESSA)(NHDOE, 2019, 2021) from the 2014-15 to the 2018-19 school year. The NHDOE officially ended the PACE innovative assessment system in March 2022. The contrasting groups standard setting method (Cizek & Bunch, 2007a) has been applied each year for five years as the local assessment information varies by year. The contrasting groups standard setting method was selected because of the design of the assessment system, which does not collect item-level data or portfolio-level data. Therefore, standard setting methods applied in previous performance assessment reform efforts as noted earlier (i.e., ID matching and Body of work) were not selected. To better contextualize how the contrasting groups standard setting method was applied and why it was selected, the

next section explains the PACE assessment system in detail.

NH PACE Assessment System

The PACE assessment system does not rely on state annual standardized achievement tests to make determinations of student proficiency in all federally required grades and subjects. Local assessment information (including performance assessment results) alongside teacher judgment is used to determine student proficiency, except in those grades and subject areas where the state achievement test is administered. There is a state-level achievement test (known as NHSAS) administered once per grade span that acts as an external audit on the system (see Table 1).

The PACE assessment system is operational in a subset of schools and districts in the state; therefore, some schools and districts administer the NHSAS in all required grades whereas PACE participating schools and districts follow the PACE model depicted below. At its largest, PACE was comprised of around 10,000 students and 13 districts; the state of New Hampshire has around 180,000 students. PACE is guided by a theory of action collaboratively developed by school, district, and state leaders (Marion & Leather, 2015).

Annual determinations of student proficiency in NH PACE schools and districts are based on local summative classroom assessment data aligned to state competencies from teacher grade books (NHDOE, 2016a). Figure 1 shows how the local summative classroom assessments (including common and local performance-based assessments) result in end of year competency scores for each student. Common assessments (known as PACE Common Performance Tasks) are performance assessments created by representatives of all participating PACE districts and administered by all participating PACE districts in every grade and subject area where there is not a state-level achievement test. The common assessments or PACE Common Tasks are used to calibrate scoring across districts and enhance the comparability of annual determinations of student proficiency (Evans & Lyons, 2017b). The student scores included in standard setting derive from two sources: district-level end of year competency scores and teacher judgment survey results. Each is discussed in more detail below.

End of Year (EOY) Competency Scores

End of year (EOY) competency scores are similar to final averaged grades. Imagine that there are ten curriculum-embedded, summative classroom assessments administered over the course of the year in grade 3 mathematics. Table 2 illustrates how the EOY competency score in this context would be the mean (or simple average) of those summative classroom assessments. The EOY competency scores are on the grading scale of the district or school and are submitted for every student in a PACE participating district or school in every federally required grade and subject area to the NH DOE at the end of each school year.

Table 1. Grades/Subjects included in the NH PACE Assessment System vs. State Standardized Assessment System

Grade	English Language Arts	Math
3	NHSAS	PACE
4	PACE	NHSAS
5-7		PACE
8	NHSAS	NHSAS
9-10	PACE (no accountability)	PACE (no accountability)
11	SAT	SAT

Figure 1. Graphical Overview of the NH PACE Assessment System

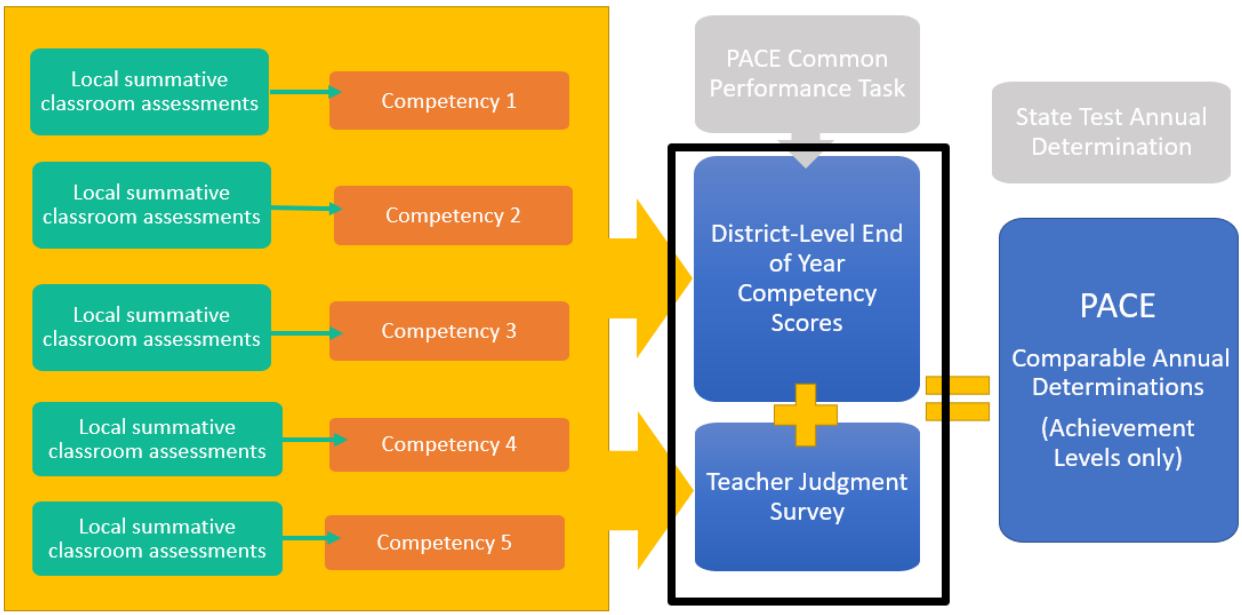


Table 2. End of Year (EOY) Competency Score Computation Example

Student ID	Summative Classroom Assessments (on the grading scale of the district)										EOY Competency Score
92789087	1.2	1.5	2.0	2.0	2.2	2.5	3.0	3.0	3.0	3.1	2.35
92789088	2.0	2.0	2.0	2.2	3.0	3.0	3.3	3.5	4.0	2.0	2.70
92789089	2.8	3.0	3.1	3.5	2.0	2.5	3.0	2.7	3.0	3.2	2.88
….											

Teacher Judgment Survey

The Teacher Judgment Survey asks classroom teachers to classify each of their students based on grade level and content-specific Achievement Level Descriptors (ALDs). ALDs articulate the expected levels of performance related to the knowledge and skills described by the grade-level content standards. The ALDs range from 1 to 4 and are aligned with the state's ALDs for the standardized assessment system—NHSAS. The levels represent the four levels of achievement that are federally reported. Level 3 is considered proficient.

Teachers are instructed to carefully read the PACE ALD for their grade/subject and then consider each student's achievement level based on their *cumulative knowledge* of each student's independently completed classroom assessments, student work and other evidence of learning. It is exactly because teachers are familiar with their students' achievement throughout the year that they are asked for their judgment of student achievement using the ALDs. The teacher judgment is intended to be a holistic judgment about students' achievement given everything they know about the student from the whole year of work relative to the expectations outlined in the ALDs. Teachers are further instructed to look for the closest match between each student's performance and the ALDs using a preponderance of evidence approach. Students do not need to meet every aspect of a descriptor, but

the teacher should use their best judgment to match them with the correct achievement level. Students well-below grade level should receive the lowest rating (level 1) and students performing above the proficiency descriptor (level 3) should receive the highest rating (level 4).

Quality of EOY Competency Scores & Teacher Judgments

It is important to recognize that the quality of standard setting assumes the quality of the local assessments that comprise the EOY competency scores, the quality of the local scoring of those assessments, and the quality and accuracy of the teacher judgments. This is especially pertinent given the known reliability issues with the use of performance assessments for large-scale accountability uses (Davey et al., 2015; National Center for Education Statistics, 1996; Tung & Stazesky, 2010) and the social moderation needed to ensure accurate teacher judgments of student achievement (Klenowski & Wyatt-Smith, 2013). However, it is beyond the scope of this article to describe the ways in which the PACE assessment program collected evidence of summative classroom assessment alignment and local assessment quality, evidence of reliable local scoring, or conducted social moderation comparability audits to show within- and cross-district comparability in expectations of student performance. More details on those evaluations can be found in the PACE Technical Manuals (Center for Assessment, 2020a, 2020b), generalizability studies (Evans & Lyons, 2017a), and in the following article that explains the comparability challenges related to PACE (Evans & Lyons, 2017b).

Standard Setting Method Explanation

The purpose of standard setting is to designate cut scores that define the four levels of achievement for the PACE Annual Determinations. Standard setting plays a central role in the validity of the interpretations from the scores (Cizek, 2011). This is especially true for PACE due to four main reasons:

1. PACE standards must be re-set every year due to differences in local classroom summative assessment information that comprise end-of-year competency scores.
2. PACE does not report out any individual-level scale scores beyond the annual determinations. This places extra burden on the validity of the interpretations drawn from the achievement level placements.
3. Each PACE district has a unique scale associated with their end of year competency scores. Even if the scales are nominally the same (e.g., 1.00-4.00) the interpretations associated with the score points will differ across districts due to differences in scoring practices. Therefore, PACE standard setting is used as a critical aspect of comparability for the PACE assessment system.
4. The PACE assessment system is required to produce annual determinations that are comparable to the statewide assessment system. Therefore, the standard setting methodology is grounded in achievement level descriptors aligned across local systems. Each of the achievement levels is intended to carry the same interpretations about what students know and can do whether they participate in PACE or NHSAS.

Over five years of standard setting, the PACE assessment system has leveraged multiple variations within a general approach and refined psychometric processes to continuously improve as the assessment system scales. The contrasting groups standard setting methodology (Cizek & Bunch, 2007b) has been the primary method applied.

Rationale for Selection of Contrasting Groups Standard Setting Methodology

There are two broad categories of standard setting methods: test-centered methods and examinee-centered methods (Jaeger, 1989). In the PACE assessment system there is only one piece of information that is based on the same assessment across all students in a given grade and subject area such as Grade 5 math—and that was one common performance assessment. Given the known person by task by occasion interactions that limit the generalizability (or reliability) of score information from any one performance assessment (Ruiz-Primo et al., 1993), a test-centered standard setting method was deemed inappropriate. However, there are two other pieces of information collected across all students in a given grade and subject area: EOY competency scores and teacher judgment survey results. Given that the EOY competency scores are based on different local

assessments and local scoring approaches, and the teacher judgment survey produces an evaluation of students relative to achievement level descriptors, an examinee-centered standard setting method was selected.

There are various examinee-centered methods for standard setting where "performances of real examinees are evaluated relative to the performance standard"—in this case achievement level descriptors (Cizek, 2011, p. 61). The contrasting groups method is an examinee-centered method where "participants categorize examinees into two groups, an upper group who have clearly met the standard, and a lower group who have not met the standard, and the score that best discriminates between these two groups is taken as the cutscore" (Cizek, 2011, p. 61). This methodology aligned well with the four groups created from the teacher judgment survey results (Levels 1-4) and allowed for three cut scores to be determined on the scale of the EOY competency scores. Federal law requires at least three achievement levels. The NH PACE assessment system and application of contrasting groups standard setting methodology is explained in more detail in the next section.

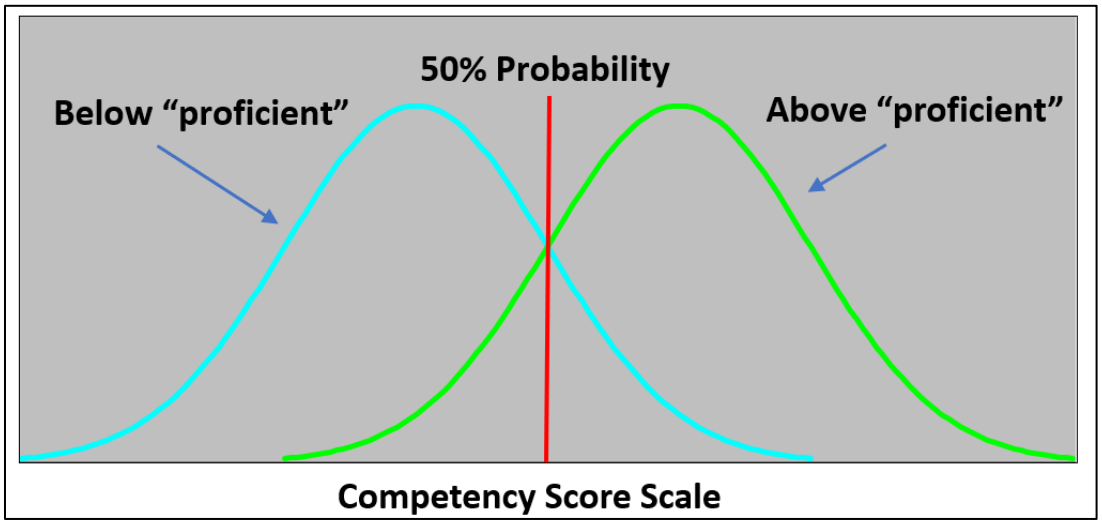
Contrasting Groups Standard Setting Methodology

The NH PACE standard setting method involves two primary steps: 1) collecting teacher judgments regarding students' perceived achievement and 2) setting cut scores on each districts' end of year competency score scale (scale refers to each district, grade, and subject combination) using the teacher judgements in a contrasting groups methodology.

Every PACE teacher completes a teacher judgment survey at the end of the school year to indicate which achievement level best describes each of their students. The subject and grade specific ALDs are entered into an online survey where teachers can easily read the descriptions and match their students to the appropriate achievement level. This process relies heavily teacher knowledge of each of their students and on a common understanding and interpretation of the ALDs.

The contrasting groups standard setting methodology involves comparing end-of-year competency scores with the teacher judgment scores to determine the cut scores that most accurately classify the students into the achievement levels. Logistic regression is used to determine the point in the score distribution where examinees have a 50% chance of being classified in the next performance level or above (e.g., the probability that a student with a score of X has a 50% or greater probability of being classified in Level 3 or higher). Figure 2 shows a graphical depiction of the contrasting groups methodology with the red line representing the cut score between the two levels. A logistic regression is estimated separately

Figure 2. Graphical Depiction of Contrasting Groups Standard Setting Methodology as Applied to NH PACE



for each cut point—Level 2, Level 3, and Level 4—in each district, subject, and grade since each local assessment system is unique. This standard setting methodology is repeated each year as the local assessment information varies from year-to-year.

Contrasting Groups Standard Setting Challenges

Standard setting, whether for standardized tests or for performance assessments systems, rarely goes as planned. There are a myriad of issues and challenges that can arise. This section explains some of the key standard setting challenges that have arisen with respect to the NH PACE assessment system.

The purpose of bringing up these challenges and the ways in which we attempted to solve those problems is to serve as an example of how one state performance assessment system anticipated likely challenges, created quality control and quality assurance measures, but then adapted over time to new and changing conditions. The full set of quality control and quality assurances processes and procedures now in place for the NH PACE system are described in the Technical Manual (Center for Assessment, 2020a). Specific challenges have been chosen that are illustrative of what might be considered likely or common challenges, especially for assessment systems design with non-standardized assessments and few common assessments shared across participating schools and districts.

Data Quality Issues Encountered: Quality Control Processes and Procedures

The first set of standard setting challenges have to do with data quality issues surrounding the end of year competency scores and teacher judgment survey results. Some of these data quality issues were anticipated in the design phase; others were not anticipated. Table 3 shows selected data quality issues and the solutions applied for the NH PACE system. The solutions applied are varied. For the first two issues, data quality control checks and district flagging business rules are used to ensure the quality of factors related to producing cut scores and are completed prior to calculating PACE cut sores. Cut score calculation rules are used with respect to the third issue to ensure consistency in setting standards by delineating rules for different scenarios. Each illustrative data quality issue is discussed in more detail below.

Data quality control checks. The first issue encountered when examining end of year competency scores was out of bound values. This data quality issue was anticipated. For example, if a district's competency score scale is 1.00 to 4.00, it is sometimes the case that data is entered that is out of that range—such as

Table 3. Selected Data Quality Issues & Solutions Applied for NH PACE Standard Setting

Selected Data Quality Issues	Our Solution
Out of bound values reported for EOY competency scores (e.g., 0.75 on a scale of 1.00-4.00)	Instituted data quality control checks
Unexpected distributions of teacher judgment survey ratings (e.g., no variance; reduced variance; bimodal distributions)	Revised PACE ALDs and provide more guidance (e.g., clarify relationship of TJS to CB grading) Created district flagging business rules that require certain actions and set *a priori* criteria
Logistic regression does not converge or other logistic regression issues (e.g., small sample sizes—one district has only 5 students per grade!)	Created additional cut score calculation business rules and document how each cut score was calculated for each scale (which rule was employed)

0.75. A series of data quality control checks was instituted to systematically examine the data quality on a series of factors prior to running the logistic regression. These data quality control checks included flagging out of bound values and viewing raw data by scale (district, grade, and subject) to complete human reasonableness checks. We use scatterplots of end of year competency scores by teacher judgment survey ratings for each district, grade, and subject combination.

District flagging business rules. Another data quality issue that was less anticipated yet has had a significant effect on the contrasting group standard setting methodology, is unexpected distributions of teacher judgment survey ratings. Unexpected distributions may be due to a number of factors, including: (a) misunderstanding or lack of clarity about the TJS process and PACE ALDs; and/or (b) unique features of participating schools or districts. With respect to the potential misunderstanding or lack of clarity about the TJS process and PACE ALDs, we instigated multiple conversations with PACE school/district leadership and teacher leaders to understand what might be unclear, confusing, or ambiguous. These conversations led to two changes. First, a rewriting of the PACE ALDs so they were more narrative descriptions yet still aligned with the NHSAS ALDs. And, second, revision to the directions related to the TJS so that the relationship between competency-based grading and TJS ratings was clarified. Conversations with PACE educators revealed erroneous beliefs about competency-based grading that were then applied to the TJS ratings. Specifically, a bias was reported about not rating students a Level 1 or Level 4 because in some schools and districts a Level 1 was perceived to be a poor reflection on the teacher and a Level 4 was incorrectly assumed to mean that the student is doing above grade-level work. Additional layers of training and resources were provided to local PACE school/district leadership and teachers to try to improve the quality of TJS ratings. For more information see the PACE Annual Performance Report to the U.S. Department of Education (NHDOE, 2019).

With respect to unique features of participating schools/district, PACE includes small rural districts as well as wealthy, high-performing districts. These two types of districts can present challenges to the contrasting groups standard setting method in particular because small rural school districts may only have 5-10 students per grade/subject area and the teacher judgment survey results may have little or reduced variance. The lack of variance results from the small sample size, not necessarily inaccurate teacher judgments—but the two factors are conflated and make running logistic regression and isolating three performance level cuts extremely difficult. Similarly, some high performing districts have reduced variance because most of their student population tends to score high on the state test and, in parallel fashion, reported TJS ratings tend to have reduced variance.

In order to examine reduced variance in the teacher judgment survey ratings, submitted teacher judgment survey ratings are analyzed by district, grade, and subject in order to identify unexpected distributions of teacher judgment prior to calculating PACE cut scores. The flagging rules evaluate variability in the teacher judgment survey ratings by district, grade, and subject in three ways: (1) identify instances where there is *no variance* in teacher judgment survey ratings (e.g., all 3s); (2) identify instances where there is *reduced variance* in teacher judgment survey ratings (e.g., all 2s and 3s); and (3) identify instances where there is *bimodal distribution* of teacher judgment survey ratings (e.g., all 1s and 3s).

Instances where teacher judgment survey ratings show evidence of no variance, reduced variance, or bimodal distribution are then analyzed using the Table 4 decision matrix below. The decision matrix guides follow-up decisions with districts and was created to balance the need for district follow-up with the realities of data issues that result from very small student populations and higher performing districts. Step 1 is a simple examination of the sample size in the district, grade, and subject combination. Step 2 is an examination of the percent of students proficient or above from prior state standardized assessment results for the district and subject in the grade level closest to the grade level under investigation. Given the design of the PACE assessment system and based on the number of years the district has been involved in PACE, the available state assessment data may be limited (e.g., grade 3 ELA, grade 4 Math, or grade 8 ELA and math).

The complete district flagging business rules analysis along with the subsequent decisions related to each flag based on the decision matrix is reported to the NH DOE by the Center for Assessment each year.

Table 4. PACE Flagging Rules for Variability in TJS Ratings Decision Matrix

Flag for TJS Ratings	Step 1: Examine Sample Size	Step 2: Examine Prior State Standardized Assessment Results
No variance	<=5 students→no follow-up >5 students→go to Step 2	Percent of students proficient is within ± 5% of the prior state standardized assessment results→no follow-up. Otherwise, the district will be contacted by the NH DOE or the Center for Assessment to verify the teacher judgment survey results.
Reduced variance	<=15 students→no follow-up >15 students→go to Step 2	

It is atypical to contact districts for follow-up based on no variance, reduced variance, or bimodal distributions in the teacher judgment survey ratings. In most years, teacher judgment survey ratings tend to concentrate in Levels 2 and 3 (about 75% of the time), the other 25% of judgments are distributed between Levels 1 and 4.

If follow-up with districts on the distribution of their teacher judgment survey ratings is deemed necessary, the business rules specify that the Center for Assessment will not calculate cut scores until teacher judgment survey results can be verified with the district. If the teacher judgment survey results cannot be verified with the district, then the district will be notified that they will receive PACE determinations for the year, but the district will need to take NHSAS along with submitting PACE data in the following year. Results from NHSAS in the following year will be compared to PACE standard setting results and if within ± 5% on percent proficient or above in the same grade and subject area then the district will not need to administer the NHSAS the following year. Otherwise, the process will continue until the district meets the ± 5% on the proficiency threshold.

Cut Score Calculation Business Rules. Another consequence of small student populations or range restriction in the end of year competency scores related to contrasting groups is that the logistic regression often does not converge or there are other logistic regression issues. While we had certain cut score calculation business rules in place during the first few years of the NH PACE standard setting, we found those business rules were not comprehensive enough, some of the methods (e.g., equipercentile linking) assume cut scores are calculated and available for other grades within the same subject area for a district which may or may not be true, and there was no systematic process for documenting what business rule was applied to calculate cut scores for every scale. Remember that a scale is every grade, subject, and grade level cut scores. The lack of documentation inhibits transparency such that external audiences are aware of what cut scores for which district, grade, and subject combinations are calculated by logistic regression and which were set using some alternative method or business rule.

To address these standard setting challenges a comprehensive set of cut score calculation business rules were created. Cut score calculation rules are used to ensure consistency in setting standards by delineating rules for the following: addressing every possible pattern of presence/absence of teacher judgments placing student achievement in each achievement level; describing the statistical process (dichotomous logistic regression) used for estimating cut scores where there are sufficient data; and ensuring consistency in calculating cut scores when there are problems with estimating a cut score using the logistic regression.

There are two major parts in cut score calculation: (1) initial cut score calculations, including logistic regression of teacher judgments of students' achievement being at or above a given achievement level on students' mean competency scores to estimate cut scores for a given scale (a scale is a district, grade, and subject combination); and (2) alternate cut score calculations for situations in which the logistic regression does not converge or in which the logistic regression found a lower probability of students being at or above a specific achievement level associated with increases in mean competency scores. The alternate cut

score calculations followed a set of business rules. For each scale with at least one cut score where the logistic regression was problematic, do the following:

1. Create a three-bit string identifying for each cut score whether the cut score calculation was problematic. For example, "011" indicates that the cut score between levels 1 and 2 was successfully calculated (=0), but the cut scores between levels 2 and 3 and levels 3 and 4 were problematic (=1).
2. Using the three-bit string identified in the prior step, follow the rules for calculation given in the corresponding row of Table 5, which shows up to three ordered calculations (i.e., first calculation, second calculation, third calculation).

The full set of business rules are detailed in the PACE Technical Manual (Center for Assessment, 2020a). Additionally, documentation is submitted each year (since standards must be reset each year) that shows the results of the contrasting groups standard setting analyses with applied cut score calculation business rules (Center for Assessment, 2020b).

Contrasting Groups Quality Assurance Challenges

Prior to submitting the calculated cut scores as final, several quality assurance impact analyses are conducted to evaluate the consistency and stability of the cut scores. The purpose of these quality assurance process and procedures is to review the outcome and reasonableness of the cut scores produced using historical data to flag results that seem unlikely or unreasonable given trends over time for each scale.

Historical data from previous years of the PACE and NHSAS system are used alongside the most recent year of data whenever possible. Impact analyses are run at both the system- and district-level. The impact analyses include:

- Cohort analysis: Examines how students in a given grade/subject perform in comparison to students in the same grade/subject for the previous year and any other years of data available using percent of students proficient or above.
- Longitudinal analysis: Compares how students in a given grade perform in the previous grades (same subject) for the previous year and any other years of data available using percent of students proficient or above.
- State test analysis: Compares proficiency rates between PACE and NHSAS in grades 3-8 using percent of students proficient or above by subject.
- Performance level analysis: Compares the percent of students in each performance level (1, 2, 3, or 4).

The difficulty encountered with the standard setting quality assurance impact analyses is how to interpret results. Specifically, what are acceptable levels of variance over time within cohorts, within districts, and/or between state assessment systems. With respect to the variance between state assessment systems, what is the arbiter of 'truth' and who gets to make that decision. How close is close enough, given that he two state assessment systems are designed with two very different theories of action and use vastly different forms of data to produce student determinations. Interpreting analyses is also made even more challenging as the number of PACE districts and schools does not stay consistent year-to-year: some join, some drop out. The purpose is to scale, which then makes comparisons and interpretations more difficult. Below we share two examples from the 2019 impact analyses to illustrate difficulties with interpretation.

Longitudinal Analyses

Figures 3 and 4 below show PACE results from the 2019 longitudinal analyses for all districts combined based on the percent of students proficient or above for the graduating class of 2024. Figure 3 shows English language arts results and Figure 4 shows mathematics. Both contain results from the 2015-16 school year to the 2018-19 school year.

The key question when using longitudinal data to provide quality assurance relative to the PACE annual determinations is what level of variance should be expected and fall within a "normal" range of variation from year-to-year for a graduation cohort by subject area? Our solution was to compare the variance

Table 5. Business rules for calculating cut scores based on whether each logistic regression had problematic results

Needed	Cut12	Cut23	Cut34
001			Cut34 <- MaxPossCS
010		Cut23 <- (Cut12 + Cut34) / 2	
011		Cut23 <- (Cut12 + MaxPossCS) / 3	Cut34 <- MaxPossCS
100	Cut12 <- (MinPossCS + Cut23) / 2		
101	Cut12 <- (MinPossCS + Cut23) / 2		Cut34 <- MaxPossCS
110	Cut12 <- (MinPossCS + MinPossCS + Cut34) / 3	Cut23 <- (MinPossCS + Cut34) / 2	
111	Cut12 <- (MinPossCS + Cut23) / 2	Cut23 <- (MinPossCS + MaxPossCS) / 2	Cut34 <- MaxPossCS

Note. MaxPosCS = scale-specific maximum possible competency score (or HOSS when HOSS = Highest Observable Scale Score); MinPosCS = scale-specific minimum possible competency score (or LOSS when LOSS = Lowest Observable Scale Score)

Figure 3. Longitudinal Analyses for All Districts Combined Based on Percent Proficient or Above for the Graduating Class of 2024 in ELA from the 2015-16 SY to the 2018-19 SY

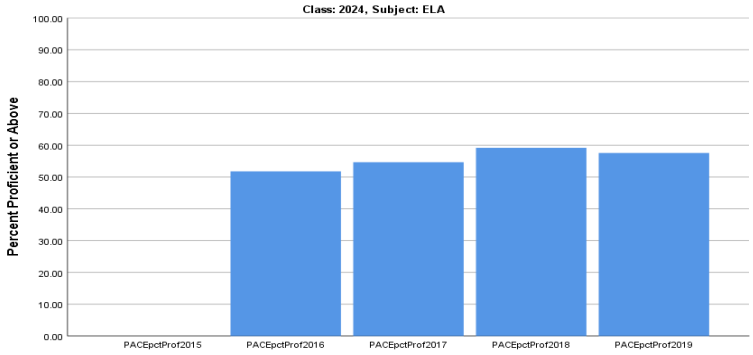
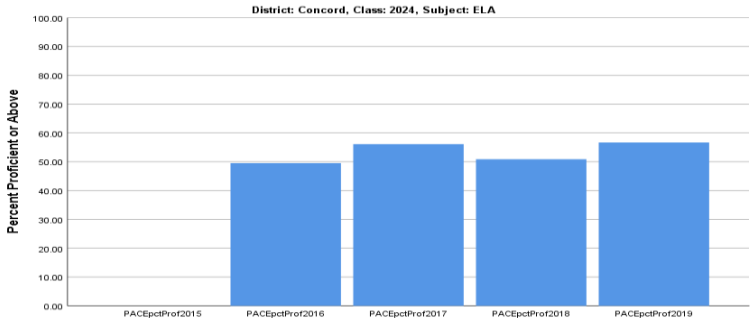


Figure 4. Longitudinal Analyses for the Concord School District Based on Percent Proficient or Above for the Graduating Class of 2024 in ELA from the 2015-16 SY to the 2018-19 SY



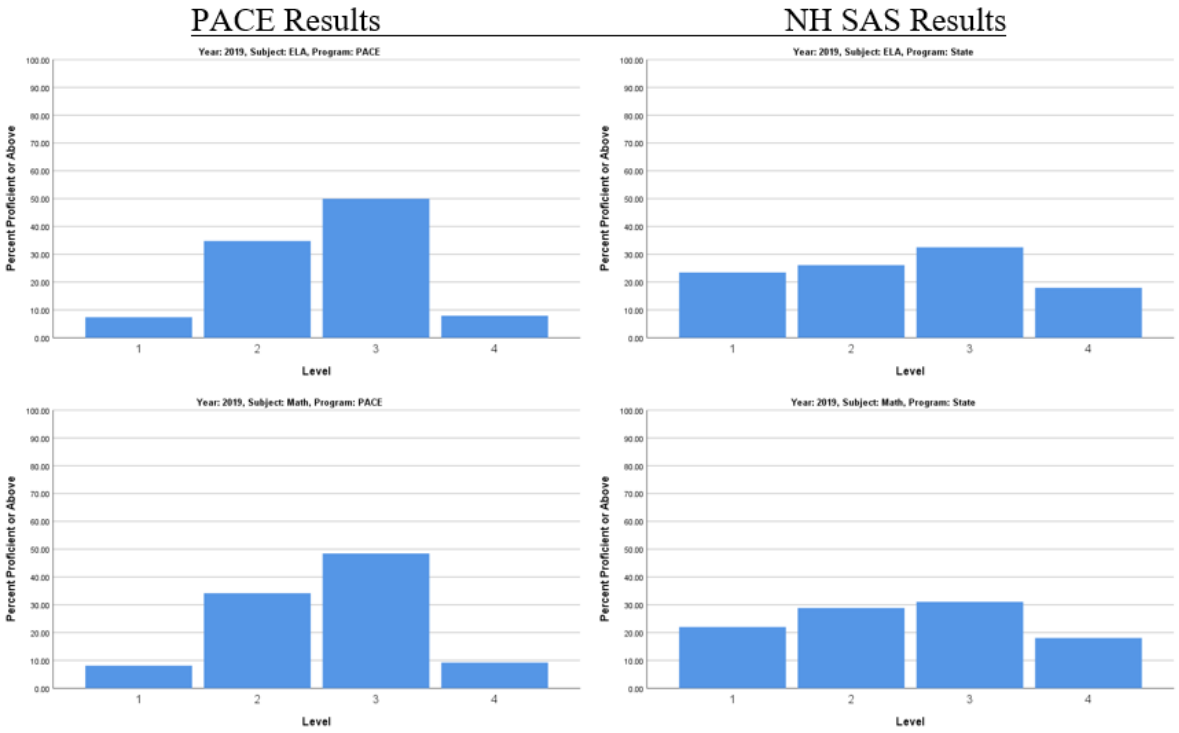
in state test results for the same graduating cohort by subject area and use that variation to set an *a priori* threshold against which we could evaluate the reasonableness of the PACE variation. These types of analyses are imperfect, however, because students move in-and-out of PACE schools and districts just like they move in-and-out of the state. How different would the percent proficient or above need to be in order to make a judgment that PACE standard setting results are not an accurate reflection of student achievement across all PACE schools and districts?

Performance Level Analyses

Figure 5 illustrates the differences in achievement levels resulting from the NH PACE versus NHSAS assessment systems. The top panel shows differences for English language arts and the bottom panel for mathematics from the 2018-19 school year. Notice that the NH PACE annual determinations show reduced variance across performance levels as a higher percentage of students are classified as Levels 2 & 3 in comparison to the more uniform distribution of NHSAS achievement levels. The NHSAS *by design* has a more even distribution across the four performance levels.

The key questions from a quality assurance perspective include: Is the reduced variance in PACE achievement levels problematic from a practical, policy, or technical perspective? Can (or should) PACE standard setting methods or results be adjusted to even out the distribution? For example, an equipercentile linking or stabilizing approach to previous NHSAS results (since once per grade span testing is used) could be applied to even out the PACE performance levels. This approach assumes the NHSAS result are the arbiter of truth and students should not change performance levels between NHSAS administration once per grade span. Are these assumptions problematic? Or, should policymakers and practitioners simply care about the Level 2 to 3 cut because that is the cut that determines proficiency. Moreover, what level of comparability is required between results either at the performance level or at the proficient/not proficient level between the two state assessment systems? These are open questions

Figure 5. NH PACE vs. NHSAS percent of students at each proficiency level for ELA (top panel) and Math (bottom panel) from the 2018-19 school year



and relate as much to policy and practical concerns, as to technical quality.

Conclusion

Large-scale performance assessment systems bring unique challenges, including unique challenges related to standard setting. The goal of this paper was to explain the standard setting methodology applied to one state assessment system and use that system to illustrate challenges, applied solutions, and lessons learned that may apply to similar types of large-scale performance assessment systems. Given previous pendulum swings with respect to performance- or portfolio-based assessment systems in the 1980s and 1990s, many of which were scrapped due to concerns about the technical quality of system results, it behooves assessment system designers to pay careful attention to the standard setting methods applied and how those methods may need to be adapted or audited to promote technical quality, stakeholder buy-in, and positive community perceptions.

References

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

Aurora Institute, Center for Assessment, Center for Innovation in Education, Envision Learning Partners, Great Schools Partnership, & KnowledgeWorks. (2021). Measuring forward: Emerging trends in K-12 assessment innovation. Retrieved from https://knowledgeworks.org/resources/emerging-trends-k12-assessment-innovation/

Center for Assessment. (2020a). New Hampshire's Innovative Assessment System. Performance Assessment of Competency Education (PACE). Evaluating Technical Quality (Volume 1): Manual. Dover, NH: Author. Retrieved from https://www.education.nh.gov/sites/g/files/ehbemt326/files/files/inline-documents/pacetechmanualvol1.pdf

Center for Assessment. (2020b). New Hampshire's Innovative Assessment System. Performance Assessment of Competency Education (PACE). Evaluating Technical Quality (Volume 2): Results. 2018-2019 School Year Edition. Retrieved from https://www.education.nh.gov/sites/g/files/ehbemt326/files/files/inline-documents/pacemanualvol2results.pdf

Cizek, G. J. (Ed.). (2011). *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). New York, NY: Routledge.

Cizek, G. J., & Bunch, M. B. (2007a). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Cizek, G. J., & Bunch, M. B. (2007b). The contrasting groups and borderline group methods. In *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications, Inc.

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

Darling-Hammond, L., & Snyder, J. (2015). Meaningful learning in a new paradigm for educational accountability: An introduction. *Education Policy Analysis Archives*, *23*(7). Retrieved from http://dx.doi.org/10.14507/epaa.v23.1982

Davey, T., Ferrara, S., Holland, P. W., Shavelson, R., Webb, N. M., & Wise, L. L. (2015). Psychometric considerations for the next generation of performance assessment. Princeton, NJ: Educational Testing Service.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*(4), 289–303.

Evans, C. M., & Lyons, S. (2017a). Application of generalizability theory to classroom assessments in a school accountability context. In *National Council for Measurement in Education*. San Antonio, TX.

Evans, C. M., & Lyons, S. (2017b). Comparability in balanced assessment systems for state accountability. *Educational Measurement: Issues and Practice*, *36*(3), 24–34. https://doi.org/http://dx.doi.org/10.1111/emip.12152

Ferrara, S., & Lewis, D. M. (2011). The item-descriptor (ID) matching method. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives (2nd ed.)*. New York, NY: Routledge.

Fine, M., & Pryiomka, K. (2020). Assessing college readiness through authentic student work: How the City University of New York and the New York Performance Standards Consortium are collaborating toward equity. Palo Alto, CA: Learning Policy Institute. Retrieved from https://learningpolicyinstitute.org/sites/default/files/product-files/RCA_CUNY_Assessing_College_Readiness_REPORT.pdf

Guha, R., Wagner, T., Darling-Hammond, L., Taylor, T., & Curtis, D. (2018). The promise of performance assessments: Innovations in high school learning and college admissions. Palo Alto, CA: Learning Policy Institute. Retrieved from https://learningpolicyinstitute.org/product/%0Apromise-performance-assessments.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd edition, pp. 485–514). New York, NY: Macmillan.

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

Kingston, N., & Tiemann, G. C. (2011). Setting performance standards on complex assessments: The body of work method. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives (2nd ed.)*. New York, NY: Routledge.

Klenowski, V., & Wyatt-Smith, C. (2013). *Assessment for education: standards, judgement, and moderation*. London: Sage.

Koretz, D. M., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, *13*(3), 5–16.

Marion, S., & Leather, P. (2015). Assessment and accountability to support meaningful learning. *Education Policy Analysis Archives*, *23*(9). Retrieved from http://dx.doi.org/10.14507/epaa.v23.1984

Massachusetts Department of Education. (2020). Application for new grants under the competitive grants for state assessment program. Retrieved from https://oese.ed.gov/files/2020/10/Massachusetts-Dept.-of-Elementary-and-Secondary-Education_Redacted.pdf

National Center for Education Statistics. (1996). *Technical Issues in Large-Scale Performance Assessment (NCES 96-802)*. (G. W. Phillips, Ed.). Washington, DC: National Center for Education Statistics, Office of Educational Research and Improvement.

New Hampshire Department of Education. (2016a). Application for inclusion in Performance Assessment for Competency Education PACE 2016-2017. Concord, NH: Author.

New Hampshire Department of Education. (2016b). Moving from good to great in New Hampshire: Performance Assessment of Competency Education (PACE). Concord, NH: Author.

New Hampshire Department of Education. (2019). NH PACE 2018-19 IADA Annual Performance Report. Retrieved from https://www.education.nh.gov/sites/g/files/ehbemt326/files/files/inline-documents/nhpaceapr1819.pdf

New Hampshire Department of Education. (2021). Performance Assessment of Competency Education. Retrieved from https://www.education.nh.gov/who-we-are/division-of-learner-support/bureau-of-instructional-support/performance-assessment-for-competency-education

Parsi, A., & Darling-Hammond, L. (2015). Performance assessments: How state policy can advance assessments for 21st century learning. National Association of State Boards of Education. Retrieved from https://www.nasbe.org/performance-assessments-how-state-policy-can-advance-assessments-for-21st-century-learning/

Pecheone, R., Kahl, S., Hamma, J., & Jaquith, A. (2010). Through a looking glass: Lessons learned and future directions for performance assessment. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement*, *30*, 41–53.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*, 215–232.

Stecher, B. (2010). Performance assessment in an era of standards-based accountability. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Tung, R., & Stazesky, P. (2010). Including performance assessments in accountability systems: A review of scale-up efforts. Boston, MA: Center for Collaborative Education.

Yen, W. M., & Ferrara, S. (1997). The Maryland School Performance Assessment Program: Performance Assessment with Psychometric Quality Suitable for High Stakes Usage. *Educational and Psychological Measurement*, *57*(1), 60–84. https://doi.org/10.1177/0013164497057001004

Citation:
Evans, C. (2023). Applying a contrasting groups standard setting methodology to a large-scale performance assessment program used for accountability. *Practical Assessment, Research, & Evaluation*, 28(4). Available online: https://scholarworks.umass.edu/pare/vol28/iss1/4/

Corresponding Author:
Carla M. Evans
National Center for the Improvement of Educational Assessment

Email: cevans [at] nciea.org