Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
xxxx 2017, Vol. 00, No. 0, pp. 1–11

Comparability in Balanced Assessment Systems for State Accountability

Carla M. Evans, *Education Department, University of New Hampshire, Durham,* and Susan Lyons*, *National Center for the Improvement of Educational Assessment, Dover*

The purpose of this study was to test methods that strengthen the comparability claims about annual determinations of student proficiency in English language arts, math, and science (Grades 3–12) in the New Hampshire Performance Assessment of Competency Education (NH PACE) pilot project. First, we examined the literature in order to define comparability outside the bounds of strict score interchangeability and explored methods for estimating comparability that support a balanced assessment system for state accountability such as the NH PACE pilot. Second, we applied two strategies—consensus scoring and a rank-ordering method—to estimate comparability in Year 1 of the NH PACE pilot based upon the expert judgment of 85 teachers using 396 student work samples. We found the methods were effective for providing evidence of comparability and also detecting when threats to comparability were present. The evidence did not indicate meaningful differences in district average scoring and therefore did not support adjustments to district-level cut scores used to create annual determinations. The article concludes with a discussion of the technical challenges and opportunities associated with innovative, balanced assessment systems in an accountability context.

Keywords: accountability, assessment system design, comparability, competency-based education, performance-based assessments

Accountability has influenced the use and design of assessments for the past two decades and pervades the current context (Hamilton, Stecher, & Klein, 2002; Hargreaves & Braun, 2013). Some have argued that the negative effects of standardized accountability tests on curriculum and instruction occur because of a fundamental misalignment between the purpose of assessment and the role assessment has played in schools (Resnick & Resnick, 1992; Shepard, 2000). This disconnect can lead to an incoherent system of assessments that do not provide instructional feedback to teachers, narrows the curriculum to focus on only those standards and subjects tested on state assessments, and drives the teaching and learning of fragmented bits of knowledge rather than deeper learning (Darling-Hammond, Wilhoit, & Pittenger, 2014; Pellegrino, Chudowsky, & Glaser, 2001; Smith & O'Day, 1991). There has been an increasing call for multiple assessments to be designed and used as a "balanced," "comprehensive," or "next generation" assessment system (Council of Chief State School Officers, 2015; Darling-Hammond et al., 2014; Heritage, 2010; Pellegrino et al., 2001; Stiggins, 2006). The challenge lies in designing assessment and accountability systems that can support instructional uses while serving accountability purposes (Baker & Gordon, 2014; Gong, 2010; Marion & Leather, 2015).

One example of using an assessment system to provide information from the classroom to the statehouse, while fulfilling federal accountability purposes, is currently taking place in New Hampshire (NH). In March 2015, the U.S. Department of Education officially approved New Hampshire's Performance Assessment of Competency Education (NH PACE) pilot project for a two-year waiver (2014–2015 and 2015–2016 school years) from federal statutory requirements related to annual state-level achievement testing (NHDOE, 2015). The NH PACE pilot was granted an additional 1-year waiver for the 2016–2017 school year.

In the NH PACE system, local assessments administered throughout the school year contribute to students' overall competency scores which are used to make annual determinations for state and federal accountability. Therefore, one key technical challenge of the NH PACE system, and likely any balanced assessment system that does not rely solely on standardized achievement tests, is using the information from multiple, local assessment sources to support comparable accountability determinations.

*The order of the authors was determined by flipping a coin. Both authors contributed equally to this article.

Carla M. Evans, Education Department, University of New Hampshire, 62 College Road, Morrill Hall 308, Durham, NH 03824; carla.m.evans@gmail.com. Susan Lyons, is an associate with the National Center for the Improvement of Educational Assessment (Center for Assessment), 31 Mount Vernon Street, Dover, NH 03820; slyons@nciea.org.

Because the NH PACE system is implemented currently with only a subset of school districts in the state, there is a need to evaluate comparability at two levels: between districts implementing the PACE system and across the two assessment systems operating within the state at once. This article reports on a study to test methods for evaluating and strengthening the comparability claims between districts within the PACE assessment system. Between-district comparability is a prerequisite for evaluating comparability between the two assessment systems (Lyons & Marion, 2016). Key considerations in the design of the methods include the scalability of the methods long-term and the suitability for the New Hampshire context.

The article is organized as follows. We first examine the literature in order to define comparability outside the bounds of strict score interchangeability and explore methods for estimating comparability that support a balanced assessment system for state accountability such as the NH PACE pilot. We then estimate comparability in the context of the first year of the NH PACE pilot. Specifically, we apply two methods for estimating comparability that are used in international contexts—one aspect of the external moderation process used in Queensland, Australia (Queenland Studies Authority [QSA], 2014) that we call consensus scoring and one type of cross-moderation used in England called rank-ordering (Bramley, 2005, 2007). We use these two methods to collect evidence about the degree of comparability of judgments about student work among PACE districts in order to strengthen claims made about the utility, validity, fairness, and accuracy of the annual determinations for use in an accountability framework. The article concludes with a discussion of the technical challenges and opportunities associated with innovative, balanced assessment systems under the Assessment and Accountability Demonstration Authority as part of the *Every Student Succeeds Act* (2015).

Background

NH PACE Pilot Project

The NH PACE pilot allows selected NH school districts to base annual determinations of student proficiency in English language arts, math, and science in Grades 3–12 on a combination of local, common, and state-level assessments (Table 1) (NHDOE, 2014). The PACE system is designed to support deeper learning for students and organizational change for schools and districts (Marion & Leather, 2015). As such, high-quality, curriculum-embedded performance assessments are the cornerstone of this new accountability model.

Comparability is particularly important in the NH PACE system because there is a need to make annual determinations from different local assessments administered by the participating districts that can be used for statewide accountability purposes. Statistical equating methods that link different tests or test forms cannot be used in this context where most of the assessment information is unique to each participating district. However, there is a set of performance assessments, referred to as PACE Common Tasks, administered in all participating districts in 17 subject and grade combinations (Table 1). The PACE Common Tasks are designed to serve as calibration tools, providing evidence about the comparability of judgments related to student achievement across NH PACE districts (Figure 1).

Defining Comparability

Similar to validity, comparability is not an attribute of a test or test form, nor is it a yes/no decision. Instead, comparability relates to the degree to which the scores resulting from different assessment conditions can support the same inferences about what students know and can do. In this way, judgments about comparability are inherently score-based. In most large-scale assessment programs, especially in the United States, evidence of comparability is typically gathered to support the interchangeability of scale scores. In the case of an innovative assessment system such as NH PACE, it is at the level of the *annual determinations* for which we want to make comparable inferences about student achievement across participating districts. This means that if a student is reported as "proficient" in one district, that student could also be expected to be rated as "proficient" had s/he presented the same evidence of proficiency in a different district. The annual determination of proficiency, therefore, carries all the same interpretations, no matter the school district or the particular assessment information that led to that determination.

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), a determination of what evidence constitutes a satisfactory degree of comparability for any type of score should be made based on clear and defensible documentation for how comparability is examined and evaluated. In other words, a "reasoned appraisal of the strength of evidence for comparability and against comparability" should be made (Winter, 2010, p. 8). Comparability is defined, therefore, as the degree to which the results of assessments intended to measure the same learning targets produce the same or similar inferences.

Comparability is important because it relates directly to the validity and fairness of large-scale assessment and

Table 1. Local, Common, and State-Level Assessments Used to Make Annual Determinations in NH's PACE Pilot Project

Grade	ELA	Math	Science
3	Smarter Balanced	*Common and Local PBAs*	Local PBAs
4	*Common and Local PBAs*	Smarter Balanced	*Common and Local PBAs*
5	*Common and Local PBAs*	*Common and Local PBAs*	Local PBAs
6	*Common and Local PBAs*	*Common and Local PBAs*	Local PBAs
7	*Common and Local PBAs*	*Common and Local PBAs*	Local PBAs
8	Smarter Balanced	Smarter Balanced	*Common and Local PBAs*
9	*Common and Local PBAs*	*Common and Local PBAs*	*Common and Local PBAs*
10	*Common and Local PBAs*	*Common and Local PBAs*	*Common and Local PBAs*
11	SAT	SAT	*Common and Local PBAs*

Note. PBAs = performance-based assessments.

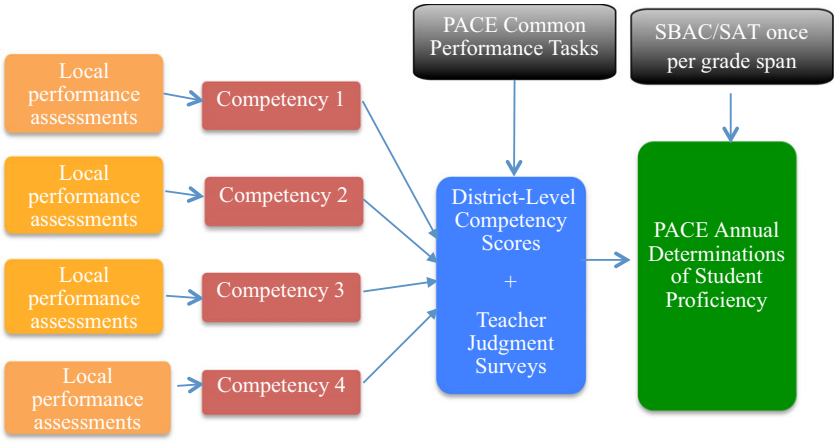

FIGURE 1. NH PACE pilot annual determination graphic. [Color figure can be viewed at wileyonlinelibrary.com]

accountability systems (AERA, APA, & NCME, 2014). In other words, can the scores resulting from different assessment conditions be used to support the same uses (e.g., school evaluation)? Comparability becomes important when we make the claim that students and schools are being held to the same standard, particularly when those designations are used in a high-stakes accountability context.

Methods to Estimate Comparability

There are many different methods for gathering evidence to support score comparability evaluations. Contrasting conceptions of comparability typically include statistical and judgmental approaches, or some combination of the two (Baird, 2007; Newton, 2010). The chosen approach is dependent upon the nature of the assessments and the intended interpretations and use of the test scores (Gong & DePascale, 2013). Some specific examples of comparability methods include: equating, calibration, projection, statistical moderation, and social moderation (Linn, 1993; Mislevy, 1992). Statistical approaches often rely on norm-referenced methods, whereas, judgmental approaches often rely on criterion- or standards-referenced methods (e.g., Sadler, 1987).

Because the NH PACE system is designed around the use of locally implemented performance-based assessments, "both the desire for comparisons across time and groups employing different performance tasks creates the demand to find a way of judging the comparability of performances on different tasks" (Linn & Baker, 1996, p. 97). Because teachers' contextualized judgments are used to score performance-based assessments in educational settings, not all methods of gathering comparability evidence are useful to support the annual determinations generated from these differing sets of performance tasks. One method that is useful, however, is social moderation. Mislevy (1992) defines social moderation as the use of "judgment to match levels of performance on different assessments directly to one another" (p. 25). In moderation, assessments are not assumed to measure the same construct, but instead judgment is used to match distributions of student performance to obtain a "correspondence table of 'comparable' scores" (Mislevy, 1992, p. 23). In some contexts such as Queensland, Australia (QSA, 2014) and in England (Newton, Baird, Goldstein, Patrick, & Tymms, 2007), social moderation has been used for decades to provide quality control for their exam systems where different tests are administered by different boards to different students with the need for comparability among scores. The method typically involves using expert judgments to audit a certain number of exams already scored by boards to evaluate the comparability of judgments across different boards and different exams within and across years (Adams, 2007; Queensland Curriculum & Assessment Authority, 2014). According to Linn (1993),

> At a minimum, information about the judgment process, the qualification of the judges, and the degree of inter-judge agreement needs to be provided. Documentation of the degree to which different groups of judges agree that given examples of performance on different tasks meet common standards also needs to be provided. Confidence in the comparability may be further buttressed by statistical comparisons made possible by the use of some common tasks. (p. 100)

In England, on the other hand, more statistical comparative judgment techniques deriving from Thurstone's (1927) *Law of Comparative Judgment* are used by employing paired comparisons or rank-ordering designs (e.g., Bramley, 2007; Pollitt & Elliott, 2003).

Alternative Methods to Estimate Comparability

There are two international examples of moderation audits that are described in this article: Queensland, Australia and England. These two examples were chosen for several reasons. First, they use moderation processes to audit teachers' summative judgments related to school-based assessments in high-stakes contexts. Second, they provide different examples of how moderation can take place. And, importantly, they have been shown to be sustainable across time. For instance, in both Queensland and England, moderation audits have taken place on senior exam systems for decades. As such, they serve as relevant, feasible, and useful exemplars for the NH PACE system. We briefly outline the strengths and challenges of these two moderation approaches and then describe the Year 1 moderation audit design for the NH PACE pilot.

Queensland, Australia: External Moderation

In 1972, Queensland replaced external exams at the end of Year 12 with a system of externally moderated school-based assessments (QSA, 2014). In this system, teachers make judgments about standards achieved by their students, including summative judgments for reporting and college admissions

purposes. In order to ensure that the levels of achievement in subjects match the requirements of syllabi, the Queensland Curriculum and Assessment Authority (QCAA) conducts a seven-phase external moderation process: syllabus development, work program approval, monitoring, verification, comparability, confirmation, and random sampling.

Queensland's external moderation process is designed to minimize key threats to comparability. For example, the seven phases of moderation standardize the content, administration, and scoring of student work, as well as the curriculum implemented. The external moderation process also supports teachers' judgments about student work and provides external advice and feedback at multiple levels about how well judgments about students' level of achievement match the performance standards.

Queensland's moderation approach requires significant infrastructure and organizational capacity to support the entire external moderation process. Seven layers of external review take place every year and those reviews are organized and findings collected and analyzed by a governing agency, the QCAA. In addition, the QCAA oversees the distribution of findings, provides professional development, and supports teachers, schools, and districts in making high-quality judgments about students' level of achievement.

One potential limitation for the U.S. context is that Queensland's external moderation process is a state-centralized operation, which intentionally limits local flexibility in curriculum, instruction, and assessment decisions in order to achieve more comparability. For example, teachers are not given flexibility to design their own syllabi, but must follow a prescribed work program and assessment design. While this may not be an issue for certain contexts that are used to high levels of state control, for other contexts with rich histories of local control and flexibility, the seven-phase external moderation process may be too restrictive.

England: Cross-Moderation

Comparability has been a fundamental concern in England's examination system since the mid-19th century because of the high-stakes nature of decisions relative to examinations given by different awarding bodies (Tattersall, 2007). Instead of folios of student work as in Queensland, ages 16+ (school leaving) and 18+ (university entry) students in England take written exams that are used in conjunction with internally assessed coursework to make university placements and job selections (Newton, 2007b). Different awarding bodies (or examination boards) create subject syllabi based on the same standards, codes of practice, and curricular structures, but each board's examination is different. This necessitates comparability of examination standards to ensure fairness, equity, and consistency.

Similar to Queensland's system of external moderation, England's system of cross-moderation has many advantages. It takes into account the demands of collecting and randomizing student work samples prior to moderation panels and uses a methodology (paired or rank-ordered comparisons) that allows judgments of relative quality to be placed on a scale. Statistical techniques used, especially multimodeling methods, allow more controlled examination and a "genuine methodological advance" over other methods (Newton, 2007a, p. 456).

However, there is contention over the use of statistical techniques in comparability studies (Goldstein, 2007; Johnson, 2007; Newton, 2007a). These concerns seem to center on the assumptions underlying statistical models and whether or not those assumptions hold in this context. For example, statistical models assume that it is possible to control for the various factors that may impact the performance standard attained, thereby interpreting differences in judgments between examining boards and reviewers to "differential grading standards" (Newton, 2007a, p. 461). But if those assumptions do not hold, invalid conclusions may result.

Application to the New Hampshire Context

The approach for estimating comparability between NH districts implementing the PACE system must be philosophically coherent with the PACE system in that it fulfills the purpose of ensuring public and political confidence in the comparability of achievement levels among districts implementing PACE. In addition, the moderation audit should fit the NH context, which is one of local control and flexibility, and be feasible to implement and sustain in the long term.

One of the motivating reasons why NH is piloting a new kind of accountability system is because the state wants to support meaningful learning and continuous improvement models, as well as promote shared accountability between districts and the state (Marion & Leather, 2015). Fostering district agency is not only philosophically coherent with a competency-based model of education that is intended to support student agency, but is also coherent with NH's long history of local control. As such, it is unlikely that adopting either Queensland's or England's system of moderation *in toto* would fit the motivation behind the PACE system or the context of NH. Therefore, adapting moderation practices from Queensland and England may be the most helpful for the NH PACE system.

NH PACE's Moderation Audit

The primary goal of NH PACE's moderation audit (or any moderation audit) is quality control: to gather evidence of the degree to which there are systematic differences in the stringency or leniency of scoring across participating districts. The theory behind NH PACE's moderation audit is that if District A, for example, consistently scores their students' PACE Common Performance Tasks more leniently than the other districts as revealed in the cross-district rescoring, it is reasonable to assume that this leniency may transfer to the scoring of students' local performance tasks and assessments in District A. Therefore, the district-level competency scores, based on locally scored tasks and assessments, may carry different inferences about student achievement than the other districts. This information could then be used to inform any decisions about adjustments to the cut scores for District A's annual determinations. In other words, the NH PACE moderation audit will be used to detect and correct threats to comparability between districts implementing PACE, and ultimately strengthens claims of comparability for the resultant annual determinations.

Beyond providing quality assurance and quality control about the comparability of annual determinations among PACE districts, the NH PACE moderation audit can also provide information to facilitate conversations about leniency or stringency of scoring at the school and district level. These conversations are particularly useful because they may not only improve teacher assessment literacy, but also improve

how teachers judge the quality of student work relative to the state competencies. In addition, the information from the moderation audits also informs ongoing task and rubric design as consistency in scoring across districts is related to the quality of the performance tasks and rubrics.

Data, Methods, and Results

Participants and Sample

During the summer of 2015, 85 teachers and leaders from seven of the eight 2015–2016 NH PACE implementing districts participated in a full-day moderation audit using student work samples from the four implementing districts in 2014–2015: District A, District B, District C, and District D (high school only). The four implementing districts in 2014–2015 were asked to select 12 student work samples for each of the 17 PACE Common Performance Tasks that both spanned all score points on the analytic rubric (1–4) and was as close as possible to representative of the underlying distribution of student achievement within their district. Districts were asked to supply 12 student work samples for each of the 17 PACE Common Performance Tasks to minimize the burden on teachers and districts, while ensuring enough student work samples to test for district, subject area, and grade level effects. The analytic rubrics were task-specific rubrics developed by cross-district teams of teachers in conjunction with the PACE Common Performance Tasks. The number of rubric dimensions ranged from 2 to 6, but all used a 4-point scale. Due to these differences in the number of dimensions by PACE Common Performance Task, we averaged across the analytic rubric scores for the purposes of these analyses. In all, 396 student work samples were gathered from the 17 PACE Common Performance Tasks: 353 student work samples were used in the consensus scoring method and 43 student work samples were used in the rank-ordering method. The 43 student work samples from high school science used in the rank-ordering method were not consensus-scored because every district gave a different performance task. In other words, there was no common performance task administered in high school science in Year 1 that could be used for consensus scoring. We discuss this issue in more detail under the rank-ordering method below.

Consensus Scoring Method

Similar to the random sampling review panels (external moderation process) in Queensland, Australia, the consensus scoring method involves pairing teachers together, each representing different districts, to score student work samples. Teachers (or other district leaders such as literacy coaches and principals) were paired based on their grade level and content area expertise with a teacher from another district with the same grade level and content area expertise. These pairs of teachers were then assigned student work samples from neither of their districts. Seventy-eight teachers and leaders from the 2015–2016 PACE implementing districts were involved in the consensus scoring method. After training and practice, both judges within each pair were asked to individually score their assigned samples of student work and record their scores. Working through the work samples one at a time, the judges would discuss their individual scores and then come to an agreement on a "consensus score." The purpose of collecting consensus score data is to approximate "true scores" for the student work, which can serve as calibration weights to detect any systematic, cross-district differences in the stringency of standards used for local scoring.

In only three out of the 353 work samples, consensus could not be reached. This indicated that raters from different districts were able to rate student work very consistently. In these three cases, an expert scorer (who did not have affiliation with any particular district) provided a score for the work sample. Each pair of teachers was asked to score six random samples of student work in the morning. In the afternoon, pairs were shuffled so that teachers were still working with colleagues representing different districts, but were assigned a new partner to score another six student work samples.

Consensus Scoring Analysis and Results

Table 2 reports the frequency of the consensus-scored student work samples by grade level, subject areas, and district.

To detect any systematic discrepancies in the relative leniency and stringency of district scoring, we averaged the analytic rubric scores across to derive a single consensus score and a single teacher score for each student work sample, and then calculated a mean deviation index. This index (Equation 1) is the mean difference between the consensus score and teacher score across all student work samples for each district as calculated by the following, for District k:

$$Deviation_k = \frac{\sum_i^n (teacher_i - consensus_i)}{n_k}. \quad (1)$$

Using this index, a negative mean deviation would indicate systematic underestimation of student scores by classroom teachers (i.e., district stringency), and positive mean deviation scores would indicate systematic overestimation of student scores by classroom teachers (i.e., district leniency). The values of the deviation metric are on the scale of the rubric points. Table 3 shows the average observed deviation

Table 2. Number of State Assigned Student Identifiers by Grade, Subject, and District

Grade	Frequency	Subject	Frequency	District	Frequency
3	26	ELA	135	District A	106
4	47	Math	156	District B	108
5	52	Science	62	District C	117
6	41			District D	22
7	58				
8	29				
9	50				
10	50				
Total	353	Total	353	Total	353

Table 3. Average Deviation by District

District	*N*	Mean Deviation	*SD*	Mean Absolute Deviation	*SD*
District A	106	.235	.66	.532	.46
District B	108	.437	.64	.614	.47
District C	117	.292	.72	.577	.52
District D	22	.394	.72	.667	.46

and standard deviation by district. As an example, the interpretation of the mean deviation for District A is that, on average, teachers in District A scored their student work on the common performance tasks .235 points higher than the same work as scored by the consensus raters. Additionally, Table 3 includes mean absolute deviations and standard deviation by district. The standard deviations associated with this metric are smaller than for the mean deviation because the possible deviation score range is constrained to positive numbers.

Across all districts, the consensus scoring yielded scores that were a bit lower than the teacher-given scores. There are a variety of hypotheses that could explain this phenomenon, but the finding itself is not necessarily problematic from a *comparability* perspective, as long as the relative leniency of the teacher-given scores is even across districts. An analysis of variance (ANOVA) was run in order to investigate the variance in the deviation index that can be attributed to differences in districts, grade level, and subject area. For this analysis, the raw deviation metric (rather than the absolute deviation) is used as the dependent variable in order to maintain the directional interpretability of the results. The factors for this three-way ANOVA are district, grade level, and subject area. The results of this ANOVA are shown in Table 4.

The results show that the variation in the deviation index across districts is statistically significant at $\alpha = .05$. However, interpretation of this finding is limited given the statistical significance of the interaction effects. The significant three-way interaction effect indicates that the relationship among district and subject area changes by grade level. This means that average deviation varies depending on the unique district, grade-level, and subject area combination.[1] Pairwise post hoc analyses (as shown in Table 5) reveal that there are no significant differences in marginal mean deviation among any two districts; rather, the significance in the district main effect is likely driven by the interaction effect (i.e., differences in the unique district, grade, and subject units).

One interesting pattern that emerges from examining grade level and subject area means for each district is that District B seems to have a higher deviation index than the other districts in many of the subject areas and grade levels. With the exception of the middle school grade levels, where the deviations in math and English language arts (ELA) are counterbalanced, District B seems to have a more lenient standard in scoring than the other districts. To test this hypothesis, we ran a follow-up ANOVA to test for a district by school type (e.g., elementary, middle, and high school) effect. The results of this ANOVA are presented in Table 6. In this model, district and school type alone are not sufficient for explaining variation in the deviation index; however, the district by school type interaction effect is statistically significant at the $\alpha = .05$ level.

Table 4. Deviation by District, Grade, and Subject—ANOVA

	df	*F*	Sig.	Effect Size $\eta^2_{partial}$
District	3	3.108	.027	.029
Grade	7	1.969	.059	.043
Subject	2	.332	.717	.002
District * Grade	14	2.201	.008	.091
District * Subject	4	2.698	.031	.034
Grade * Subject	4	.655	.623	.008
District * Grade * Subject	8	2.078	.038	.051

Table 5. Follow-Up Pairwise Comparisons

District 1	District 2	Mean Difference	Standard Error	Sig.
District A	District B	−.2018	.08670	.124
District A	District C	−.0565	.08503	1.000
District A	District D	−.1590	.14857	1.000
District B	District C	.1452	.08462	.523
District B	District D	.0428	.14833	1.000
District C	District D	−.1024	.14736	1.000

Table 6. Deviation by District and School Type—ANOVA

	df	*F*	Sig.	Effect Size $\eta^2_{partial}$
District	3	2.112	.098	.018
School Type	2	1.149	.318	.007
District * School Type	4	3.527	.008	.040

In sum, the consensus scoring approach required assigning pairs of raters from two different districts to review samples of student work and to assign a consensus score to that piece of work. These consensus scores were used as an anchor for comparing the locally assigned scores from the different districts. The results suggest that though the differences are small in effect size, there remains a need for additional training on scoring and within-district calibration, as well as for increased cross-district calibration. There was not enough evidence based on these results to make any cut score modifications this year. Had there been enough evidence of systematic scoring differences in a particular grade level and subject area within a district we would have applied an equipercentile standard setting procedure.

Rank-Ordering Method: High School Science

High school science presented a special challenge in calibrating the cross-district scores because there were no PACE Common Tasks across districts; each district assigned completely unique tasks for the three subject areas—earth science, physical science, and life science. Typically, score calibration procedures require one of two conditions to be met: (1) common persons across tasks, or (2) common tasks across persons. Because neither of these conditions was satisfied in the 2014–2015 implementation of PACE Common Tasks for high school science, we looked to alternate methods of score calibration. Therefore, the high school life science calibration process for PACE was modeled after the rank-ordering cross-moderation method used in England.

Unlike the consensus scoring methodology used for the other PACE subjects and grade levels, the calibration method used for high school science involved an individual rank-ordering process. The seven participating judges were given packets of student work that had been grouped by average rubric score, and were asked to rank-order the student work based on quality, or, in other words, evidence of achievement in science. Each packet contained 10 student work samples and student work from all four districts was represented in each packet. The order of the papers placed in the packet was arbitrary within average rubric scores. Each teacher was asked to rank four packets, which ensured that every teacher

saw 40 of the 43 student work samples—12, 10, 10, and 11 from District A, District B, District C, and District D, respectively.

Judge Training

Before beginning the ranking exercise, judges were first asked to familiarize themselves with each of the different tasks. In order to do so, the judges read through blank copies of the tasks and the associated task description, teacher directions, and student instructions. Then, for each task, a district representative was asked to briefly provide an overview of the performance task, including any parts that were particularly useful for discriminating among students and items or parts that were particularly difficult or did not run smoothly (because this was often the first year the tasks has been implemented). Judges then took the opportunity to ask clarifying and follow-up questions of the district representative. There were no high school science teachers present from District C, so in order to familiarize themselves with the task from that district, judges discussed their impressions of the task in pairs, which was followed with a large-group discussion of the task.

Once the judges felt comfortable with the four tasks, judges were trained on the ranking process. The instructions for the judges were based also on similar studies completed in England. As Bramley (2007) explains, "The need for the whole exercise in first place arises from the fact that the different boards (districts in the case of PACE) have different specifications and question papers. The judges are really therefore being asked to judge which performance is better, taking into account any differences in the perceived demands of the questions (and specifications)" (p. 265). The judges involved in the PACE calibration for high school science were likewise instructed to rank papers based on merit, evidence of student understanding, demonstrated competence, and student knowledge of the nature of science, which are all different ways of saying "better," as Bramley (2007) succinctly puts it. In order to minimize construct-irrelevant variance, judges were also explicitly told not to rank on such qualities as handwriting, grammar (neither of which was relevant to construct), length, and the quality of the copy.

Rank-Ordering Analysis and Results

The data sets resulting from the rank-ordering method were reorganized to represent dependent[2] pairwise comparisons. The Thurstone model for paired comparison data was used to fit a unidimensional scale representing quality, on which each student work sample was placed (Equation 2):

$$\ln\left[P_{ij}/\left(1-P_{ij}\right)\right] = B_i - B_j, \qquad (2)$$

where P_{ij} is the probability that work i beats work j, and B_i and B_j are their respective estimates on the unidimensional scale. Similar to the discrepancy analysis completed for the consensus score results, the Thurstone scores can be compared to the original teacher scores with a deviation index. However, unlike the consensus scores, the Thurstone scores are not on the same scale as the local, teacher-given scores. To account for the differences in the scales, both sets of scores were first transformed into standard scores before calculating the deviation index (Equation 3):

$$Deviation_k = \frac{\sum_i^n \left(zscore_{\text{teacher}} - zscore_{\text{Thurstone}}\right)}{n_k}. \qquad (3)$$

In this equation, the Thurstone scores are treated as our best estimate of the true quality of the student work. The deviation metric can be interpreted similarly to the deviations calculated from the consensus scores, where positive deviations indicate district leniency (i.e., the teacher-given scores are higher than the estimated quality measure). These deviation metrics differ in that the units are not on the scale of the rubric scores, but rather represent standard deviation units. The Thurstone model ran successfully and has good data-model fit indices. There were no student work samples or raters with Infit greater than 2.0, and only one paper with Outfit greater than 2.0. Additionally, the separation reliability is .96, indicating that the rank-ordering procedure resulted in a strong differentiation in quality for student work. Both distributions of scores are approximately normal and there is a moderate linear relationship between the Thurstone measure and the rubric scores ($r = .526, p < .001$). The existence of a linear relationship means the scores on the two scales can be meaningfully compared and we can infer that the judgmental ranking analysis yielded reasonable estimates of student work quality.

In order to determine if any of the districts are scoring systematically more leniently or stringently than the other districts, a deviation analysis was run. Table 7 shows the average deviation score for each district.

These mean differences indicate systematic differences in the location of district work in the rubric score and Thurstone score distributions. Though district designation is conflated with other factors, including task and student population, because all tasks were judged by the same people and placed on the same scale this discrepancy score represents only systematic differences in the inferences that rubric scores carry across districts, rather than differences in student populations. Unlike the results for the consensus scoring, these results cannot be directly interpreted in the scale of the rubric scores. Rather, these reflect relative differences in leniency or stringency across districts. Because judges were only asked to rank student work rather than score student work, the results of this analysis can only reveal *relative* differences in district leniency; the mean deviation metric is a "zero sum game."

The results do reveal scoring differences across districts, most notably in District D. On average, District D teachers were scoring their student work a full standard deviation below (more rigorous) where the judges placed the same student work within the sample. To a lesser extent, the opposite is the case for District C, where the rubric scores tended to be systematically higher than their rank order would suggest during the calibration. This may not necessarily mean that teachers in District C were more lenient than others, but it could be evidence that the task given in District C did not elicit evidence of achievement at the same level of rigor as the other tasks.

Table 7. Mean Deviation Scores by District

District	*N*	Mean Deviation
District A	12	.258
District B	10	.165
District C	10	.710
District D	11	−1.07

Rater bias. Because District D fared so favorably in the rank-ordering exercise, we decided it would be worth checking into the possible effects of any kind of rater bias, especially because judges participating in the rank ordering were not evenly distributed across districts. In fact, there were three judges representing District D, while all other districts had no more than one representative. One possible reason why District D seemed to fare especially well in the rank order exercise might be that judges would tend to favor the task and student work coming from their own district.

In order to search for this kind of bias, we examined the relative rank ordering of districts (across packets) by judges. Because the packets were grouped roughly by average rubric score, the quality of the work is naturally controlled for when examining the median rank of each district across packets. We did not find any evidence to suggest that judges tended to rank work from their own district more favorably than work from other districts. Rather, student work from District D was consistently ranked highly across all judges. Interestingly, the median rank orders have a high degree of spread, which indicates that the rank ordering of work within packets was strongly predicted by district, which provides further evidence to suggest there are systematic differences in the quality of work, receiving similar scores, across districts. In other words, there are cross-district differences in the inferences that can be drawn about what students know and can do from the scores in the differing districts. One likely explanation for this finding, in the case of high school science, is that the task employed in District D called for more complex scientific thinking than any of the other district tasks, which was suggested by the judges themselves in a postranking discussion. This may mean that the results of methodology are more reflective of the differences in rigor of the tasks than of any differences in teachers' abilities to score consistently. As a direct result of this rank-ordering methodology, the teachers decided to use the task from District D as the common task, across all districts, for the following academic year. In all, we take the results of the life science audit to be further evidence of the presence of the three-way interaction effect between district, subject, and grade level, but in this case it seems clear that the quality of the task has a strong influence on the quality of student work elicited.

Discussion and Conclusion

The purpose of this study was to test methods for evaluating the comparability of annual determinations between districts implementing the NH PACE system that fits the NH context and is sustainable. We explored alternative methods to estimate comparability from Queensland, Australia and England. We then applied one aspect of Queensland's seven-phase external moderation process, which we call consensus scoring, and one type of England's cross-moderation process called rank-ordering to estimate comparability in the context of the first year of the NH PACE pilot (2014–2015 school year). The assumption underlying this study is that the degree to which teachers in different districts score similarly or differently is a good window into the degree to which teachers in the various districts hold students to similar expectations on local assessments. This evidence about the degree of difference in scoring among PACE districts can then inform decisions about adjusting performance standards to reflect district differences and strengthen the claims of comparability between districts in the annual determinations.

We found that applying the two methods in the context of the NH PACE system highlighted some strengths and limitations of the two methods. First, both methods do provide comparability evidence in local scoring within a district. If results of the consensus scoring method or rank-ordering method indicate incomparability of local scoring for a particular grade level and subject area within a district, that evidence can be used alongside other sources of evidence as a rationale to make cut score adjustments. Both methods also involve teachers from multiple districts reviewing student work samples from other districts, which has the added benefit of providing a rich context for professional development. In previous research on the effects of high-stakes performance-based assessment systems on student performance (Borko, Elliott, & Uchiyama, 2002; Lane, Parke, & Stone, 2002; Parke, Lane, & Stone, 2006), professional development had a strong mediating effect on the relationship between the performance-based assessment system and changes in teacher instructional practices. Because resources for professional development at scale plagued previous large-scale performance assessment systems (Tung & Stazesky, 2010), using a social moderation comparability method not only provides the evidence necessary of comparability in local scoring, but it also provides a built-in professional development opportunity for teachers. The feedback from teachers who participated in the PACE Summer Institute (where the consensus scoring and rank-ordering took place) was overwhelmingly positive. There were many comments from teachers on the evaluation form about the ways in which evaluating student work from other districts and discussing student work with teachers from other districts was useful to their professional practice.

That said, reviewing student work samples across districts is costly and time-consuming. The practicality and feasibility of scaling up the proposed methods in a large-scale performance assessment program is a real concern particularly within a state that has many more districts or other units with a large number of different local assessment systems. One way New Hampshire has attempted to address scale issues is through improved technology. For example, in the first year of the PACE pilot, the four participating districts provided a paper copy of the requested 12 student work samples for each of the 17 PACE Common Performance Tasks. Due to copy quality, some of these student work samples could not be used and each student work sample had to be copied so that each judge had his/her own copy. The logistics of copying student work samples was not feasible after Year 1 of the pilot. Additionally, judges recorded their consensus scores or rank orders on a "scoring" sheet provided before entering their judgments into a preloaded Excel document. This process was also not scalable because errors entering student work identification numbers and scores were prevalent. In 2016, Year 2 of the PACE pilot, the project scaled up to eight school districts and 18 student work samples per PACE Common Performance Task (around 1,400 student work samples). To facilitate the almost quadruple growth in student work samples collected, New Hampshire made technological improvements by taking advantage of digital scanning and the state Learning Management System. Each district was asked to scan and then upload their student work samples. The entire consensus scoring process was completed online,

including viewing student work samples and entering scores. As this project continues to scale, New Hampshire is undergoing an intensive research and development process to procure additional software that will support the scaling of this effort.

In general, we found the consensus scoring method to be the most feasible in terms of the number of teachers required. For example, the consensus scoring method only requires pairs of teachers, which is efficient for consensus-scoring a relatively large number of student work samples over the course of the day (around 20 student work samples depending on the grade level and subject area). Since we requested a limited number of student work samples per PACE Common Task (12 in Year 1), scoring all the submitted student work samples in one day was possible with two to three teachers attending from each participating district per PACE Common Task. This number of scorers seems reasonable at scale, especially if regional scoring sessions could be held across the state to reduce/eliminate travel costs.

In contrast, the rank-ordering method requires a larger number of teachers because of the significant overlap in packets needed to estimate a Thurstone scale—around 8 teachers per 40 student work samples. The main advantage of the rank-ordering method is that it can be used when there is no common performance task administered by all the participating districts. This situation was encountered in the first year of the PACE pilot in science, but is no longer the case—there is one PACE Common Performance Task now administered in all 17 grade level and subject area combinations in which there is no state-level achievement test administered.

In terms of the evidence provided by the two methods applied in this study, due to the significance of the interaction effect between district, grade level, and subject area it is clear that further efforts to strengthen the comparability of cross-district scoring are needed. Part of this can be accomplished through higher quality task and rubric design, an effort already under way, as well as cross-district scorer training. Such comparability challenges are not unexpected during the first year of this complex pilot, and the results of this study point out areas where improvement is necessary. Taking all of the evidence into account, we did not recommend that any unilateral adjustments be made to the district-level cut scores in Year 1 of the pilot because there was not enough evidence of incomparability of local scoring within any district. In future years, if there are multiple sources of incomparability, we will likely consider adjusting particular grade-level and subject area performance standards within districts using an equipercentile standard-setting procedure.

It is important to note that there are a variety of sources of evidence gathered to support the comparability of annual determinations between districts produced from the NH PACE pilot. In addition to the evidence resulting from the moderation audits described in this article, the growing body of comparability evidence that is actively being gathered includes the development of common achievement level descriptors, estimates of interrater reliability, maps of local assessment content coverage, a body-of-work performance standards validation method, and comparisons to external assessments (i.e., Smarter Balanced and SAT). These additional sources of evidence are not described in detail in this article; however, we want to emphasize that, while the social moderation methods employed in Year 1 of this pilot are central to our comparability argument, they are by no means the only source of evidence used to support this argument.

In conclusion, our ability to implement balanced assessment systems, or assessment and accountability systems that support meaningful learning, continuous improvement, and local decision making, relies on the technical quality of the scores resulting from school-based assessments, particularly performance-based assessments. One of the major obstacles to scaling up the use of local performance-based assessments for accountability purposes is the concern that doing so would be at the expense of comparable achievement determinations across districts. This is a significant concern given the equity and fairness issues at stake. Not only must our thinking be flexible to solve this problem, but our methods must adapt to meet the measurement challenge (Lyons & Marion, 2016).

This reality highlights the competing purposes inherent in an assessment and accountability system where tensions exist between the requirements for each purpose. Although maximizing strict score comparability may be ideal for accountability purposes (i.e., cross-district and cross-state comparisons), it may be detrimental to the purpose of supporting meaningful learning and continuous improvement—a critical purpose of educational assessment. The NH PACE system attempts to balance these two distinct purposes and fulfill the needs and requirements by exploring: (1) how to achieve comparability and (2) how to determine the level of comparability that is necessary and desirable.

The recently passed (2015) *Every Student Succeeds Act* includes a provision for up to seven states to apply for an Innovative Assessment and Accountability Demonstration Authority allowing the state to pilot systems in a subset of school districts that may include:

(1) competency-based assessments, instructionally embedded assessments, interim assessments, cumulative year end assessments, or performance-based assessments that combine into an annual summative determination for a student, which may be administered through computer adaptive assessments; and
(2) assessments that validate when students are ready to demonstrate mastery or proficiency and allow for differentiated student support based on individual learning needs (Sec. 1204, p. 84).

States awarded flexibility under the Demonstration Authority will have to demonstrate that all students are exposed to high-quality instruction, have the same opportunity to learn, and are held to the same performance expectations. Comparability, therefore, must always be set in a context: comparable for what purpose, comparable at what level, and comparable to what degree? In so doing, accountability systems based on school-based assessments or other innovative assessment systems permitted under the Demonstration Authority must provide evidence to support comparability claims. Judgmental approaches to linking educational assessments, such as social moderation, are able to support such comparability evaluations, as we have shown in this article. Although comparability claims in general, and especially in an alternative accountability model based on school-based assessments, can never be unequivocal, the methods presented in this article provide tools to strengthen the body of evidence related to the comparability of scores between districts implementing the PACE system. This level of comparability is a prerequisite to providing evidence of score comparability across the two

assessment systems operating within the state at once (Lyons & Marion, 2016). As innovative accountability systems begin to emerge under the Demonstration Authority of the *Every Student Succeeds Act*, the results from New Hampshire may provide a basis for thinking about comparability in a broader, more flexible, and ultimately more useful way.

Acknowledgments

We would like to thank Dr. Scott Marion for his leadership and commitment to this project and Dr. Lorrie Shepard for her feedback on an earlier draft of the manuscript. Additionally, we would like to extend our gratitude to the New Hampshire Department of Education and the participating districts for supporting this work.

Notes

[1]Because interaction effects can be an artifact of outliers, the analysis was rerun without the most extreme cases and the results were replicated.

[2]Though the comparisons are dependent in that they are self-consistent, treatment of the comparisons as independent should produce measurements that are close to linearly related to the measures produced had the dependence been accounted for (Smith & Smith, 2007). Due to the increased computational load produced when dependent rankings are long, the pairs are treated as independent. The range of the measures will likely be overestimated and the standard errors underestimated; therefore, the results of the analysis will be treated as an upper bound on the amount of discrepancy in scoring across districts.

References

Adams, R. (2007). Cross-moderation methods. In P. E. Newton, J. A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 212–245). London, UK: Qualifications and Curriculum Authority.

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

Baird, J. A. (2007). Alternative conceptions of comparability. In P. Newton, J. A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 124–165). London, UK: Qualifications and Curriculum Authority. Retrieved from http://hdl.handle.net/1983/1004

Baker, E. L., & Gordon, E. W. (2014). From the assessment OF education to the assessment FOR education: Policy and futures. *Teachers College Record*, *116*(11), 1–24.

Borko, H., Elliott, R., & Uchiyama, K. (2002). Professional development: A key to Kentucky's educational reform effort. *Teaching and Teacher Education*, *18*, 969–987. https://doi.org/10.1016/S0742-051X(02)00054-9

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement*, *6*, 202–223.

Bramley, T. (2007). Paired comparison methods. In P. E. Newton, J. A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–300). London, UK: Qualifications and Curriculum Authority.

Council of Chief State School Officers. (2015). *Comprehensive statewide assessment systems: A framework for the role of state education agency in improving quality and reducing burden*. Washington, DC: Author.

Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for college and career readiness: Developing a new paradigm. *Education Policy Analysis Archives*, *22*(86). https://doi.org/10.14507/epaa.v22n86.2014

Every Student Succeeds Act. (2015). Pub.L. 114-95 § 114 Stat. 1177.

Goldstein, H. (2007). Commentary on statistical issues arising from chapters. In P. E. Newton, J. A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 445–447). London, UK: Qualifications and Curriculum Authority.

Gong, B. (2010). *Using balanced assessment systems to improve student learning and school capacity: An introduction*. Washington, DC: Council of Chief State School Officers and Rennaissance Learning.

Gong, B., & DePascale, C. (2013). *Different but the same: Assessment "comparability" in the era of the Common Core State Standards*. Washington, DC: The Council of Chief State School Officers.

Hamilton, L. S., Stecher, B., & Klein, S. P. (2002). *Making sense of test-based accountability in education*. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/monograph_reports/MR1554.html

Hargreaves, A., & Braun, H. (2013). *Data-driven improvement and accountability*. Boulder, CO: National Education Policy Center.

Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity*? Washington, DC: Council of Chief State School Officers.

Johnson, S. (2007). Commentary on judgmental methods. In P. E. Newton, J. A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 295–299). London, UK: Qualifications and Curriculum Authority.

Lane, S., Parke, C. S., & Stone, C. A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment*, *8*, 279–315. https://doi.org/10.1207/S15326977EA0804

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*, 83–102.

Linn, R. L., & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities. Ninety-Fifth Yearbook of the National Society for the Study of Education, Part I* (pp. 84–103). Chicago, IL: University of Chicago Press.

Lyons, S., & Marion, S. (2016). *Comparability options for states applying for the Innovative Assessment and Accountability Demonstration Authority: Comments submitted to the United States Department of Education regarding proposed ESSA regulations*. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from www.nciea.org

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

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Washington, DC: Office of Educational Research and Improvement.

New Hampshire Department of Education (NHDOE). (2014). *New Hampshire Performance Assessment of Competency Education: An accountability pilot proposal to the United States Department of Education*. Concord, NH: Author.

New Hampshire Department of Education (NHDOE). (2015, March 5). Press Release: Governor Hassan, Department of Education Announce Federal Approval of New Hampshire's Pilot. Retrieved May 22, 2015, from http://education.nh.gov/news/pace.htm

Newton, P. E. (2007a). Comparability monitoring: Progress report. In P. E. Newton, J. A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 452–476). London, UK: Qualifications and Curriculum Authority.

Newton, P. E. (2007b). Contextualising the comparability of examination standards. In P. Newton, J. A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 9–42). London, UK: Qualifications and Curriculum Authority.

Newton, P. E. (2010). Contrasting conceptions of comparability. *Research Papers in Education*, *25*(3), 285–292. https://doi.org/10.1080/02671522.2010.498144

Newton, P. E., Baird, J. A., Goldstein, H., Patrick, H., & Tymms, P. (Eds.). (2007). *Techniques for monitoring the comparability of examination standards*. London, UK: Qualifications and Curriculum Authority.

Parke, C. S., Lane, S., & Stone, C. A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation*, *12*, 239–269. https://doi.org/10.1080/13803610600696957

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

Pollitt, A., & Elliott, G. (2003). *Monitoring and investigating comparability: A proper role for human judgment*. Cambridge, UK: University of Cambridge Local Examinations Syndicate.

Queensland Curriculum and Assessment Authority (QCAA). (2014). *Moderation handbook for Authority subjects*. South Brisbane, Australia: The State of Queensland (Queensland Curriculum and Assessment Authority). Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Moderation+handbook+for+Authority+subjects#0\nhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Moderation+handbook+for+Authority+Subjects#0

Queenland Studies Authority (QSA). (2014). *School-based assessment: The Queensland assessment*. South Brisbane, Australia: The State of Queensland (Queensland Studies Authority).

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. O'Connor (Eds.), *Evaluation in education and human services* (pp. 37–75). Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/978-94-011-2968-8_3

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, *13*(2), 191–209. https://doi.org/10.1080/0305498870130207

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

Smith, E. V., & Smith, R. M. (2007). *Rasch measurement: Advanced and specialized applications*. Maple Grove, MN: JAM Press.

Smith, M. S., & O'Day, J. (1991). Putting the pieces together: Systemic school reform. *CPRE Policy Briefs*, *6*(4), 1–10.

Stiggins, R. (2006). *Balanced assessment systems: Redefining excellence in assessment*. Portland, OR: Educational Testing Service.

Tattersall, K. (2007). A brief history of policies, practices and issues relating to comparability. In P. Newton, J. A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 43–96). London, UK: Qualifications and Curriculum Authority.

Thurstone, L. L. (1927). Law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

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

Winter, P. C. (2010). *Evaluating the comparability of scores from achievement test variations*. Washington, DC: Council of Chief State School Officers.