Evaluating Comparability in the Scoring of Performance Assessments for Accountability Purposes

by Susan Lyons and Carla Evans

Researchers report on their evaluation of comparability claims in local scoring of performance assessments across districts participating in New Hampshire's Performance Assessment of Competency Education pilot project.


Susan Lyons


Carla Evans

This brief summarizes "Comparability in Balanced Assessment Systems for State Accountability," published in *Educational Measurement: Issues and Practice* (Evans & Lyons, 2017). This study evaluated comparability claims in local scoring of performance assessments across districts participating in New Hampshire's Performance Assessment of Competency Education (PACE) pilot project.

With the passage of the Every Student Succeeds Act (ESSA), there has been increasing attention on how states could design innovative assessment and accountability systems to submit for approval under the law. 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). New Hampshire's PACE project is an example of one such system. Within PACE, local and common performance assessments administered throughout the school year contribute to students' overall competency scores, which are in turn used to make annual determinations of student proficiency for state and federal

Appears in...

Issue 47
Performance Assessment: A Deeper Look at Practice and Research

This issue is an online supplement to VUE 46, which addressed the topic of performance assessment – a personalized and rigorous alternative to standardized testing that allows teachers to build on individual students' strengths and foster more equitable learning outcomes.

VUE 47 adds additional current materials, offers opportunities for additional voices, and provides more examples of performance assessment. Because performance assessment is an active national conversation, the work continues; following VUE 46's publication, important national conferences and other milestones occurred that we're able to share here. This issue also provides perspectives from students, educators, researchers, and policymakers.

Download issue as PDF

Purchase print copies

accountability. The challenge is using the information from multiple, local assessment sources to support comparable scoring across districts.

Comparability

Comparability is not an attribute of a test or test form, nor is it a yes/no decision. Instead, comparability is the degree to which scores resulting from different assessment conditions can support the same inferences about what students know and can do. 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 AND RESULTS

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). We apply two judgmental methods for estimating comparability that are used in international contexts: consensus scoring and rank-ordering.[1]

Consensus Scoring

The consensus scoring method involves pairing teachers together, each representing different districts, to score student work samples from students outside of either of their districts. The student work samples are common performance tasks given across districts in particular grade or subject areas. Examining the work samples one at a time, the judges 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. To detect any systematic discrepancies in the relatively leniency and stringency of district scoring, we calculated averages differences in local teacher scores and the consensus scores (mean deviation index). Using this index, a negative deviation indicates an underestimation of student scores by classroom teachers (i.e., district stringency), and positive mean deviation indicates overestimation of student scores by classroom teachers (i.e., district leniency).

Across all districts, the consensus scoring yielded scores that were positive, meaning they were a bit lower than the teacher-given scores. This finding itself is not necessarily problematic from a comparability

Related content

Performance-Based Assessment: Meeting the Needs of Diverse Learners

Performance Assessment Examples from the Quality Performance Assessment Network

The Future Is Performance Assessment

Seizing the Opportunity for Performance Assessment: Resources and State Perspectives

Case Study: The New York Performance Standards Consortium

Search VUE

Search for...

Search

perspective, as long as the relative leniency of the teacher-given scores is even across districts. Analysis revealed uneven scoring across districts, suggesting that there remains a need for additional training on scoring and within-district calibration, as well as for increased cross-district calibration.

Rank-Order Method

High school biology presented a unique challenge in calibrating the cross-district scores because there was no common performance assessment administered across districts in this discipline; each district developed and implemented completely unique tasks. Typically, score calibration procedures require one of two conditions: 1) common persons across tasks, or 2) common tasks across persons. Because neither of these conditions was satisfied in the 2014-15 implementation of high school science in PACE, we looked to alternate methods of score calibration and modeled our method after the rank-ordering cross-moderation method used in England.

The seven participating judges, all high school science teachers, were given packets of student work that had been grouped by average rubric score and represented student work from biology performance tasks from all four districts. After training and an opportunity to familiarize themselves with the different performance assessments from the four districts, the judges were instructed to rank papers within each packet based on merit, evidence of student understanding, demonstrated competence, and student knowledge of science, which are all different ways of saying “better,” as Bramley (2007) succinctly puts it. The rank orders from teacher judges were converted to scores, which were compared to original teacher scores.

The results revealed scoring differences across districts, most notably in one district in which teachers were scoring their student work a full standard deviation below (more rigorous) where the judges placed the same student work within the sample.

CONCLUSION

We found that applying the two methods in the context of the PACE system highlighted some strengths and limitations of each method. First, both methods provide comparability evidence in local scoring within a district. 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. Using a one of these methods not only provides the evidence necessary of comparability in local

scoring, but also provides a built-in professional development opportunity for teachers.

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. 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 through virtual task design and scoring.

States awarded flexibility under ESSA's Innovative Assessment Demonstration Authority will have to demonstrate that all students have the same opportunity to learn and are held to the same performance expectations. In so doing, accountability systems based on school-based assessments or other innovative assessment systems permitted under the Innovative Assessment Demonstration Authority must provide evidence to support comparability claims. The methods presented in this brief provide tools to strengthen the body of evidence related to the comparability of scores across districts implementing an innovative system of assessments.

Related topics: • Assessment

Display Footnotes

Display References