Psychometric Considerations for Performance Assessment With Implications for Policy and Practice

by Charlene G. Tucker, Ed.D.
Measurement Consultant for the K–12 Center at ETS

Based on the white paper
Psychometric Considerations for the Next Generation of Performance Assessment
by
Tim Davey (Educational Testing Service)
Steve Ferrara (Pearson)
Paul W. Holland (Emeritus, University of California, Berkeley)
Rich Shavelson, Chair (Emeritus, Stanford University)
Noreen M. Webb (University of California, Los Angeles)
Lauress L. Wise (Human Resources Research Organization)

I. Introduction

Performance assessments are used to measure performance in education, work, and everyday life. Perhaps the public's most commonly experienced performance assessment is the driver's examination—a combination of multiple-choice[1] questions probing knowledge of driving laws (etc.) and performance tasks measuring actual driving under real-world conditions. In education, content standards for guiding K-12 education systems have been revised recently to better support the preparation of our students for the current and future expectations of post-secondary college and career. States are likewise collaborating to develop and implement common assessment systems with a corresponding focus on critical-thinking, problem-solving, and analytical skills, a focus that is increasingly bringing performance assessment into mainstream K-12 educational assessment.

While a primary focus of this report is the integration of performance assessment into K-12 educational assessment programs, the measurement concepts explored speak to other fields, including licensure and certification, civilian or military jobs, and performance in sports. Research findings and illustrative examples throughout the report are therefore drawn from education and beyond.

The moment is ripe with promise as technology continues to enhance the possibility and practicality of various assessment approaches, including performance assessment.

The moment is ripe with promise as technology continues to enhance the possibility and practicality of various assessment approaches, including performance assessment. This said, performance assessment introduces some old and some new challenges for psychometrics.

The notion of performance assessment: Appropriate application and challenges

Performance assessment has long been a critical component in demonstrating one's ability to drive a car, skate competitively, or diagnose a medical condition. Often, the mention of performance assessment, especially in education, gives rise to an image of long, extended tasks taking hours if not days, weeks, or even months. However, this is not always the case, as performance assessments come in all shapes and sizes. Some performance tasks are short in duration, taking only a few minutes to administer (e.g., a history performance task that involves identifying the source of a painting and determining whether drawing conclusions from it about a historical event is warranted; Breakstone, Smith, & Wineburg, 2013). Others are moderate in duration, taking perhaps 15–20 minutes (e.g., a science investigation; Shavelson, Baxter, & Pine, 1991). Still others are quite extended (e.g., a term research paper). Often, short- and moderate-duration tasks are mixed within the same assessment. As a rule, the nature of the tasks and the time required for administration depends largely on what is being measured.

In the driving example, much of the *knowledge* required to drive safely (e.g., rules of the road, the meaning of signs) can be assessed effectively and efficiently using a multiple-choice test.

[1] *Multiple-choice* is used throughout this paper as the contrast to performance assessment. Most of the references to multiple-choice items or tests would also hold for other selected-response item types and for some short constructed-response item types. Multiple-choice is the most common and is therefore the example used throughout.

Because multiple-choice items are relatively quick to administer and inexpensive to score, a considerable sample of the required skills and knowledge can be assessed in a reasonable amount of time at reasonable cost. Knowledge about rules of the road might be likened to the knowledge expected of students about English grammar and conventions. This knowledge also can be assessed more effectively and efficiently through a multiple-choice test than through a writing performance assessment.

However, one cannot be confident that an individual can *drive* based on assessment only of discrete knowledge. Multiple-choice test items cannot capture the *doing* of driving, and fellow motorists would likely agree that doing is a critical skill to be assessed. Measurement experts use the word *construct* when they talk about the "concept or characteristic that a test is designed to measure" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 11), and performance assessment is a better suited approach to measuring some constructs. Ensuring that an assessment measures the construct of interest to support an intended interpretation is what measurement experts think of as *validity*.[2] It is the primary reason that performance assessment matters.

Ensuring that an assessment measures the construct of interest to support an intended interpretation is what measurement experts think of as validity. It is the primary reason that performance assessment matters.

A performance assessment is designed to measure abilities that multiple-choice questions cannot measure. Performance assessment, in the driving example, allows the drivers to demonstrate that they can use the discrete knowledge in conjunction with psychomotor and higher order problem-solving and decision-making skills in a real-world situation. Likewise, one needs to know that students can use their knowledge about language, as well as higher order thinking, planning, and organizational skills, to accomplish a real-world purpose such as writing an op-ed article or a persuasive essay.

Another reason that performance assessment matters is that it sends a signal to educators and students about what is important and expected. Aspiring drivers practice driving, in part, because that is a component of the required assessment for licensing. Likewise, the inclusion of writing performance in an assessment system signals to teachers and students the importance of the practice of writing in the curriculum.

Performance assessment, however, has its limitations. Because it is time-consuming and more expensive than multiple-choice testing, it is not possible to collect the large number of observations that is possible with multiple-choice items. Skaters have a limited number of opportunities to demonstrate their ability at the Olympics, and those few performances may or may not capture the ability they would demonstrate across a very large number of opportunities. Likewise, given the required time and cost, it may be difficult in summative educational assessments to include a large enough number of performance tasks to reliably measure student

[2] *Validity* refers to the weight and quality of evidence that supports the proposed interpretation of test scores. If an algebra test is intended to measure competence in algebra (i.e., be interpreted as measuring algebra), validity refers to the extent to which there is conceptual, logical, and empirical evidence to support the claim or proposed interpretation. In simple words, validity is the weight of evidence that supports the claim that the test measures what it is intended to measure.

ability. The *reliability*[3] or consistency of scores that measurement experts seek is enhanced by greater numbers of observations (e.g., test items).

A related challenge for all types of assessment is that of including a sufficient number of items or tasks with sufficient *variation* to reflect the full scope of the construct being measured. Given the limited sample of behavior possible due to time constraints, this is a particularly acute challenge with performance assessments. The driving performance assessment certifies that under a single set of circumstances, the driver was able to demonstrate proficiency. There are, of course, many variations in the circumstances that the licensed driver will subsequently face. Different types of weather, different types of cars, different traffic conditions, and different unforeseen events all change the demands on the driver. There will also be variation based on which people are rating the driver (e.g., how strictly examiners interpret the rubric, their mood on that day, their perception of the candidate). Yet, a driver is licensed based on one set or *sample* of circumstances because it would be much too expensive and time-consuming to design a testing program that offers opportunities to perform under all possible circumstances. Measurement experts think of this as *generalizability*.[4]

The psychometric challenges related to performance assessment are sometimes like the proverbial elephant in the room—the elephant is there, but no one wants to acknowledge or talk about it. The challenge, simply put, is that performance assessment involves complex, real-life-like tasks and parallel real-life responses that can be intricate and lengthy and limited in number due to time and cost. Measurement experts think of this as the trade-off between reliability and validity.

Measurement experts think of this as the trade-off between reliability and validity.

Genesis of two papers

With leadership and support from the Center for K–12 Assessment & Performance Management at ETS, a study group of six psychometricians worked over the past year to explore the psychometric challenges and opportunities presented by performance assessment, especially its integration into mainstream K-12 assessment systems. The work of the study group was shaped by input from external experts, stakeholder groups (state and consortia), and multiple layers of review.

The work focused on the psychometric considerations that must be addressed in the realm of *summative* large-scale assessment that may be used for accountability purposes with potential stakes for students, teachers, and/or schools. While performance assessment also may be used *formatively,* this application is not the focus of the study group's work as it does not require the

[3] *Reliability* is a measure of the *consistency of scores* on an assessment. Reliability ranges from 0 (random responding) to 1 (perfect consistency) and typically reliabilities 0.80 and above are considered desirable. The consistency might be measured from test takers' performance on the same test given on two occasions (test-retest reliability), from equivalent forms of the same test (equivalent or parallel forms reliability), or from their responses to each of the items on a single test (internal consistency reliability).

[4] *Generalizability* is an extension of reliability to cover complex measurement situations. Generalizability is a measure of the consistency of scores on an assessment simultaneously taking into account, for example, test takers' consistency in responding to the test items, consistency from one occasion to the next, consistency over raters, and the like. A generalizability coefficient ranges from 0 to 1 and is analogous to the reliability coefficient. It is particularly useful in assessing the consistency of scores from complex performance assessments that typically involve multiple raters, occasions, and tasks/items.

same psychometric scrutiny as the summative use case. In the formative use case, teachers have the opportunity to consider additional confirmatory information in their interpretation of the student's performance. In the summative use case, where there are potential consequences attached to the inferences made about a student's performance, the validity of those inferences is critical.

The study group's work resulted in two papers:

- The first paper, *Psychometric Considerations for the Next Generation of Performance Assessment* (Davey et al., 2015), referenced here as *the white paper*, is relatively lengthy and technical. Its target audience of test development and measurement experts in state departments of education, in professional associations, and in testing companies is assumed to have a certain degree of sophistication with psychometric concepts. The full white paper can be downloaded from http://www.k12center.org/publications/all.html

- The current paper is the second companion paper, *Psychometric Considerations for Performance Assessment With Implications for Policy and Practice.* It references itself simply as the *policy brief*. The policy brief provides a short translation of the white paper for a wide audience interested in the use of performance assessment on a large scale. This policy brief also can be downloaded from http://www.k12center.org/publications/all.html

Organization of the policy brief

This policy brief is framed around five questions that mirror the chapters of the white paper:

- What is performance assessment?
  (based on Chapter II: "Definition of Performance Assessment," primarily authored by Steve Ferrara)

- What are the psychometric challenges when using performance assessment for individuals?
  (based on Chapter III: "Performance Assessment: Comparability of Individual-Level Scores," primarily authored by Tim Davey)

- What are the psychometric challenges when using performance assessment in groups?
  (based on Chapter IV: "Performance Assessment: Reliability and Comparability of Groupwork Scores," primarily authored by Noreen Webb)

- How can psychometrics help capture information about complex performance?
  (based on Chapter V: "Modeling, Dimensionality, and Weighting," primarily authored by Lauress Wise)

- What overall advice does the study group offer to policymakers and practitioners?
  Throughout the policy brief, illustrative examples are provided. In some cases, the examples are familiar to those working in educational settings, and, where helpful, examples outside education are used.

II. What is performance assessment?

The study group had considerable conversation about the definition of performance assessment and which assessment types are in scope and which types are outside of that definition. While

the term is used differently in different contexts, the study group had to agree on the definition for purposes of its work.

For purposes of the white paper and the policy brief, performance assessment is defined as:

> An assessment activity or set of activities that requires test takers, individually or in groups, to generate products or performances in response to a complex task. These products or performances provide observable or inferable evidence of the test taker's knowledge, skills, abilities, and higher order thinking skills in an academic content domain, in a professional discipline, or on the job.

In the white paper, performance assessment is further described in terms of five characteristics which, taken together, distinguish it from multiple-choice types of assessment. Those five characteristics are:

- *Task*—the ways the prompt stimulates test takers or groups of test takers to respond (e.g., designing and conducting a mini-investigation to determine which paper towel holds the most water and which the least)
- *Response*—the kinds of responses required in the real-world (e.g., prepare a science fair project)
- *Scoring*—the ways in which test takers' responses are scored (e.g., multitrait rubrics to score a response several times)
- *Fidelity*—the accuracy with which tasks/responses emulate a real-world context
- *Connectedness*—the interconnectedness of the tasks/items within the assessment (e.g., a set of items that may include several short constructed-response items along with a more extensive essay related to the same problem)

III. What are the psychometric challenges when using performance assessment for individuals?

Of considerable concern in large-scale assessment is the *comparability* of individual-level scores across different parallel forms of an assessment, particularly when those test forms include performance assessment. For summative educational assessment programs, multiple test forms are often required because a single test form will not remain secure under repeated administration in an environment where stakes are attached to score use. Considerable effort is required to ensure that scores from one form can be compared to scores from another form. Thinking of an individual student, one needs to be confident that the student would receive more or less the same score regardless of the test form administered.

Comparability is an important concept for educational assessment because of the need to answer questions such as, How did Sam's performance compare to the performance of other fourth graders in the nation who took different forms of the test?

Comparability is an important concept for educational assessment because of the need to answer questions such as, How did Sam's performance compare to the performance of other fourth graders in the nation who took different forms of the test? One can only answer questions such as this if different students' scores produced in different places at different times on different test forms are directly comparable.

Measurement experts think of comparability in several ways. First, do the test forms measure the same content (i.e., substantive equivalence)? Second, if administered to the same group of students, do the two test forms yield roughly the same scores for individuals (reliability)? Third, do the two forms yield scores on the same scale (i.e., statistically equated)? Rigorous test design frameworks are needed to attain both substantive and statistical equivalence.

These conditions are easier to fulfill for assessments that rely on multiple-choice item types. Test developers can make use of large numbers of test items and still have a test that can be administered within a reasonable time. Well-established *equating*[5] and *linking*[6] procedures can be used to support the comparability of scores on tests with sufficient numbers of items.

Performance assessments add layers of complexity that psychometric methods developed for multiple-choice testing are not always well prepared to address. For most of the points that follow, the impact on score comparability would be reduced by increasing the number of tasks, which in turn would increase the investment required in terms of testing time and cost.

- Performance assessments are even more highly susceptible to security breaches under repeated use because performance tasks are more memorable than multiple-choice items.
- Test developers cannot select as large a sample of performance tasks as they can with multiple-choice items given the investment in testing time and cost that would be required.
- Some performance tasks are easier for some students, and other tasks are easier for others—a phenomenon common to all types of test items. This is called a *student-by-task interaction*, and its impact on comparability can be alleviated most readily by the inclusion of more tasks.
- Student performance on the same or similar tasks from one time to another can vary depending on the approach the student takes to completing the task. This is called a *student-by-occasion interaction*. Again, its impact on comparability would be alleviated by the inclusion of more occasions of performance assessment.
- Scoring methods for performance assessments are complex. In most cases, the judgment of human raters, supported by diligent training and monitoring, is required.

We can identify a set of possible strategies for coping with the comparability challenges associated with performance assessment:

- When possible, utilize established equating and/or linking procedures to support score comparability.

[5] *Equating* is a technical procedure or process conducted to establish comparable scores, with equivalent meaning, on different forms of the same test; it allows the test forms to be used interchangeably (Ryan, 2011).
[6] *Linking* is the practice of pairing or matching scores on two test forms of a test with no strong claim that the paired scores have the same substantial meaning (Ryan, 2011).

- Extend the test length to include more performance tasks.
- Include a greater number of shorter performance tasks instead of a smaller number of longer performance tasks.
- Augment performance assessments with multiple-choice test items linked to the performance assessment construct/task setting.
- Carefully standardize tasks and test forms.
- Accept that scores are not fully comparable and limit inferences accordingly.
- Report only group-level scores.

IV. What are the psychometric challenges when using performance assessment in groups?

People seldom perform alone; typically they work with others in teams. The capacity to do so is an important 21st century skill required for success in school, on the job, and in life. Hence we consider the use of performance assessment with groups of test takers. To reiterate, in a real-world context, which performance assessment aims to emulate, the application of knowledge and skills to create a product or solve a problem often occurs in the company of others.

…in a real-world context, which performance assessment aims to emulate, the application of knowledge and skills to create a product or solve a problem often occurs in the company of others.

Current standards that define what students should understand and be able to do to be prepared for college and careers in the 21st century, for example, address the importance of communicating and collaborating with others. Specifically, in education, from the Common Core State Standards:

- The English language arts & literacy in history/social studies, science, and technical subjects standards include a strand called speaking & listening, which calls for students to have opportunities to participate in conversations with large and small groups, to learn how to work together, and to develop useful oral communication and interpersonal skills (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010a).
- The mathematics standards stress communicative competence and specifically highlight the need for students to be able to justify their ideas and communicate them to others and to respond to others' arguments (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010b).

Beyond K-12 education, employers view the ability to collaborate with others to accomplish tasks to be a core 21st century competency that is very important for both landing and keeping a job. "Numerous reports from higher

…employers view the ability to collaborate with others to accomplish tasks to be a core 21st century competency that is very important for both landing and keeping a job.

education, the business community, and labor market researchers alike argue that such skills are valued by employers, critical for success in higher education, and underrepresented in today's high school graduates" (National Research Council, 2011, p. 1).

Because the ability to work in groups matters in school and beyond and because what gets assessed gets attention, groupwork has a place in performance assessment. Groups working on common tasks, however, present measurement challenges beyond those already discussed for individual-level performance assessment.

First, in measuring the work of students in groups, thought must be given to what outcome is being attributed to whom:

- If the objective is to determine an *individual's* ability, for example, to design and conduct a science experiment (or perform an ice-skating routine), involving others would be inappropriate and would invalidate any inference about the individual's ability. (Note: It may be appropriate to use groups to introduce students to the performance task and help them process information provided before engaging in the actual assessment activity individually.)

- If the objective, however, is to determine the *group's* ability to collectively design and conduct an experiment (or perform as an ice-skating pair), then the performance would be assessed for the group as a whole.

- *Individual* assessment would be appropriately conducted in the context of a *group* in cases where the individuals are being assessed on their *collaboration skills* within a group (e.g., ability to cooperate, to make a contribution, to listen to others' ideas, to manage conflict).

For the second scenario (assessment of a group's performance) or the third scenario assessment of an individual's ability to perform within a group), there are some special measurement considerations, as either performance can be influenced by several factors other than the skills of interest:

- *Group composition*: Who is in the group has been shown to have a marked influence on the behavior and on the learning of individuals (Webb, 1995).

- *Role assignment*: If students are assigned specific roles in groupwork (e.g., leader, listener), or even if the roles just develop naturally, the role students have in the group influences their performance within the group.

- *Type of task*: Different types of tasks can lead to different types of student involvement (e.g., more or less equally distributed participation).

- *Task*: Even when tasks are of similar type, particular tasks may work better for some groups or some individuals within groups.

- *Occasion*: Student behavior within a group (e.g., level of help-seeking) varies from one occasion to another.

- *Rating*: Student results are influenced by factors related to the rating process: the type of ratings used (e.g., the presence, frequency, or quality of a behavior), the type of rater (e.g., self, peer, expert), and rater training.

Theoretically, a student's group skills should be assessed across many performance tasks of varied types. The composition of the groups should also be varied, as well as the student's role within the groups. Finally, the rater or raters should be well-trained and certified. However, this could require an extensive number of group performance tasks and be as unpractical as testing a driver under the full range of circumstances that the driver will be licensed to navigate.

The following are some possible ways to control the influence of factors that are not related to the skills (constructs) of interest:

- To the extent practical, each student should participate in multiple group performance tasks, each in a different randomly formed group of peers.
- The role of technology continues to evolve and the use of avatars may have the potential to standardize the behavior of the other(s) in the group across test takers and across occasions.
- The focus might be shifted from the group skills of an individual to the group skills of the students in a classroom or school. Aggregating individual-level and small-group scores to higher levels may produce more dependable measures of the groupwork skills.
- Matrix sampling, or the assignment of different tasks to different groups of students, can effectively allow for a wide range of conditions to be represented without individual students needing to participate in a large number of tasks if inference is to the classroom.

V. How can psychometrics help capture information about complex performance?

Up to this point, this policy brief has focused on items that produce observable responses characterized numerically (e.g., 0, 1, or 1–6). To create measurement scales, psychometricians depend on models that represent test takers' performance as an overall score or a category (i.e., pass-fail) that translates test takers' responses into their standing on the construct(s) of interest. Performance assessment presents challenges for those models.

Standard psychometric models were developed for multiple-choice assessments, with many discrete test items that are scored as either right or wrong (i.e., scored dichotomously) and that are organized into tests designed to measure one clearly defined construct. Performance assessment is complex and not so easily modeled, making it more difficult to confidently associate test-taker performance with a score or a performance category.

Standard psychometric models were developed for multiple-choice assessments…

Some of the challenges associated with modeling performance assessment are:

- *Limited number of observations*: Psychometric models work best when there are many observations of the same construct (e.g., many questions to assess reading comprehension). The time and cost associated with performance assessment puts a practical limitation on the number of observations that are possible.
- *Complex and varied score scales*: Performance assessments are not generally scored simply as either right or wrong. They might be scored using a rubric, or multiple rubrics, on scales that range from 0–3, or 1–6, or any other variant (percentages, error rate). They may also

be scored using more unusual scales, such as the time required for a test taker to respond or some other process indicators. Further, the same performance assessment may result in multiple scores of different types.

- *Human influence (raters)*: Performance assessments are often scored by human judgment. That is, raters are trained to read or observe student work and evaluate it based on the defined scoring criteria (e.g., rubrics). While a high-level of training and monitoring greatly helps to ensure rater accuracy, rater variation can introduce measurement error.

- *Human influence (group members)*: Human influence can be particularly bothersome when the assessment is conducted in the context of groups. A student's performance on a groupwork skill (e.g., the ability to consider the ideas of others) is likely to be influenced by the behavior of the others in the group.

- *Connectedness*: The tools that psychometricians use to convert test-taker performance to a score or category work best when various test questions/activities are unconnected (i.e., they satisfy the assumption of local independence). A performance assessment typically includes a set of activities, products, and item types that are designed to be connected. A complex performance assessment task, for example, that requires a medical student to collect information and make a diagnosis may result in multiple scores based on many decisions or processes, but they would all be related to the same patient situation.

- *Dimensionality*: Most measurement models designed to estimate scores assume the test measures one construct (i.e., assumption of unidimensionality), so that the interpretation of the resulting score or performance category is clear. A performance assessment that requires a student to solve a complex multistep mathematics problem and then write about that process measures the student's ability to perform multiple skills and therefore could lead to confusion in the interpretation of the results.

A number of approaches are available to determine both how well a psychometric model fits data and how to analyze the dimensions that are impacting student scores. This said, psychometricians may need to step outside their comfort zone to explore the best ways to model the complex data that result from performance assessment.

…psychometricians may need to step outside their comfort zone to explore the best ways to model the complex data that result from performance assessment.

In making decisions about how assessment data will be modeled and translated into scores or categories, it is very important to be clear about what one wants to measure. This becomes important when making decisions about whether and how to combine data from different parts of an assessment (e.g., multiple-choice portion and performance assessment portion).

- If one is interested in capturing what the performance assessment measures (or the multiple things that the performance assessment measures) and if one does not want to lose that in the mix with multiple-choice items, a profile of scores on separate dimensions might serve that purpose.

- If, however, one is interested in making important decisions based on a single score that combines performance assessment and multiple-choice item types, one will have to consider various strategies for combining and/or weighting the data that reflect priorities and values.
- If the combination of performance and multiple-choice items are thought of as assessing a single construct of interest, the items could be scaled together as a single assessment. In that case, the many multiple-choice assessment items would outweigh the few performance assessment scores, producing a highly reliable measure of the single construct that is influenced little by the performance assessment.
- If, however, the objective is to influence educators to pay more attention to the skills that are measured by the performance assessment, a policy-weighting strategy could be employed. The multiple-choice items would be scaled separately and then the two components would be combined according to some prescribed, policy-relevant weights. Using that approach, it is possible that some precision in the overall score is sacrificed.

VI. What overall advice can the study group offer to policymakers and practitioners?

This policy brief concludes with the following advice:

1. Measure what matters. If the expectations for students include the ability to apply knowledge to solve complex problems or deliver complex performances, work with your assessment advisors to explore the options for how to best measure those expectations. What gets assessed often commands greater attention.
2. If measurement experts or members of your technical advisory committee are concerned about the emphasis on performance assessment, listen to them. They are trying to advise you responsibly.
3. In cases where the construct of interest can be measured by multiple-choice assessment items, embrace that option. These items are both practical and reliable: (a) the format is familiar, allowing students to focus on the challenge of the content; (b) the items are administered in essentially the same way to each student on each occasion and produce comparable scores in either paper-and-pencil or computer-delivered mode; (c) the items are quick to administer, allowing many observations in a given amount of time; and (d) the items can be scored accurately at very little cost in relatively little time.
4. Use complex performance assessment in cases where the construct of interest is important and cannot be assessed using multiple-choice items. Carefully define the construct(s) and select an assessment approach that requires students to demonstrate that construct in a way that is observable in their work. The necessary investment of dollars and time will need a clear justification.
5. Understand that measuring a complex construct accurately may require multiple observations, possibly under multiple conditions. Where consequences are attached to the assessment, it is critical that the inferences made are defensible. Given realistic time limitations, you may have to compromise on the length/complexity of the tasks in order to provide an adequate number of tasks under the required conditions.

6. If inferences are not needed at an individual student level, but rather at the level of the classroom or school, consider using a matrix sampling approach in which different students respond to different tasks. Increasing the number and variability of tasks may allow one to represent the construct of interest more thoroughly within reasonable bounds of time and expense.

7. Consider ways to enhance the performance assessment by combining it with other types of test items. This can be done with a mix of longer and shorter performance tasks, as well as multiple-choice items, all assessing parts of the construct of interest in the situated context of the performance assessment. Using certain psychometric techniques (e.g., generalizability theory), it is possible to determine the number of tasks, multiple-choice questions, and raters needed to produce reliable scores.

8. Consider ways of making performance assessment part of the fabric of, for example, classroom instruction. If it is not taking time away from instruction, the investment of time for extensive or multiple performance tasks will not be as much of an issue. However, using the results in a summative fashion with stakes attached will require attention to standardization of conditions; steps need to be taken to ensure that the results represent the work of the student(s) being credited.

9. Consider various ways of combining results of performance assessment with multiple-choice assessment. Scaling a few (or a single) performance task(s) together with many multiple-choice items is likely to lead to a highly reliable overall score that is impacted very little by the performance assessment. Treating the performance assessment(s) as separate from the multiple-choice assessment items and then combining them according to a policy-based weighting design will help make performance assessment scores count enough to warrant attention, but there may be a cost to the precision of the overall score.

10. Consider using performance assessment in the context of groups. Real-world performance often happens in the company of others, and the preparation of students for that real-world challenge is getting increased attention. Keep in mind, however, that performance assessment in the context of groups is particularly complex, whether the work of the group as a whole is being assessed or the work of individuals within the group. Be careful about any inferences about individual students.

A final note

The psychometric group found that performance assessment poses interesting and important challenges that provide a rich agenda for research. In the foreseeable future, technological advancements may create new solutions as well as new challenges. Given new applications of performance assessment, including common assessments of K-12 students across multiple states, vast amount of response data will be created, providing fertile ground for addressing that research agenda.

References

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

Breakstone, J., Smith, M., & Wineburg, S. (2013). Beyond the bubble: New history/social studies assessments for the Common Core. *Phi Delta Kappan, 94*(5), 53–57.

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

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010a). *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects.* Washington, DC: Author.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010b). *Common Core State Standards for mathematics*. Washington, DC: Author.

National Research Council. (2011). *Assessing 21st century skills.* Washington, DC: National Academies Press.

Ryan, J. (2011). *A practitioner's introduction to equating.* Washington, DC: Council of Chief State School Officers.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, *4*(4), 347–362.

Webb, N. M. (1995). Group collaboration in assessment: Multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis, 17*, 239–261.

For additional resources about next-generation assessments and the six multi-state assessment consortia, visit

www.k12center.org

Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management at ETS has been given the directive to serve as a catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

K-12 Center at ETS