Establishing an evidence-based validity argument for performance assessment

White Paper

Emily R. Lai
Hua Wei
Erika L. Hall
Dennis Fulkerson

September 2012
About Pearson
Pearson, the global leader in education and education technology, provides innovative print and digital education materials for preK through college, student information systems and learning management systems, teacher licensure testing, teacher professional development, career certification programs, and testing and assessment products that set the standard for the industry. Pearson’s other primary businesses include the Financial Times Group and the Penguin Group. For more information about the Assessment & Information group of Pearson, visit http://www.pearsonassessments.com/.

About Pearson’s White Papers
Pearson’s white paper series shares the perspectives of our assessment experts on selected topics of interest to educators, researchers, policy makers and other stakeholders involved in assessment and instruction. Pearson’s publications in .pdf format may be obtained at: http://www.pearsonassessments.com/research.
Abstract

Although performance assessments are not new, recent initiatives have proposed to use performance tasks in ambitious new ways, including monitoring student growth and evaluating teacher effectiveness. Given these innovative proposals, it is important to pursue a research agenda for systematically investigating (1) the validity of proposed uses and interpretations of test results and (2) the reliability of scores and decisions. First, we discuss dimensions of test score use that are important to consider when planning a validity research agenda. Next, we offer a framework for collecting and organizing validity evidence over time, which includes five important sources of validity evidence: test content, examinee response processes, internal test structure, external relationships, and consequences of test use. We then discuss scoring issues and the reliability of performance assessment results. Finally, we explore how disparate sources of evidence can be integrated into an overall validity argument.

Keywords: performance assessment, validity, reliability
Establishing an Evidence-Based Validity Argument for Performance Assessment

Recent years have seen a resurgence in the popularity of performance assessment (PA). A very general definition of a PA is an assessment in which the examinee is required to demonstrate his or her knowledge or skill. Examinees may be asked to physically perform the skill, where the performance itself is observed or recorded for evaluation, as when people applying for a driver’s license are required to drive a car. Examinees may also be asked to create some sort of product that can provide evidence about the targeted knowledge and skills, as when candidates for the bar are required to draft a legal brief or memo during the bar exam. Certain types of PAs allow one to evaluate processes as well as end products (e.g., a science laboratory task, in which the appropriateness of lab procedures can be assessed in addition to the accuracy or correctness of conclusions drawn). In addition, certain types of skills (e.g., the ability to plan and compose during writing) may be easier to assess via PAs than with more traditional item types (such as selected-response items). According to Lane & Stone (2006), well-designed PAs tend to exhibit certain attributes relative to more traditional assessment approaches, including (1) greater directness of measurement or fidelity to criterion performances (i.e., what some call “authenticity”) and (2) transparency or meaningfulness to examinees.

Recent policy initiatives have contributed to conversations about attributes of PAs. For example, according to the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium, performance tasks must “measure rigorous content and students’ ability to apply that content.” Such tasks will “elicit complex demonstrations of learning and measure the full range of knowledge and skills necessary to succeed in college and 21st-century careers.” Additionally, these tasks will “send a strong, clear signal to educators about the kinds of instruction and types of performances needed for students to demonstrate college and career readiness” (PARCC, 2010, p. 35). Clearly, performance tasks are distinguished from selected-response items in that examinees are producing or
creating their own response for the former but not the latter. Less clearly, performance
tasks appear to be distinguished from constructed response items by their greater
complexity and length.

Notwithstanding ambiguity in the definition of PAs, the validity of any assessment
must be evaluated with respect to proposed purposes, uses, and interpretations. Although
PAs themselves are not new, recent policy initiatives have focused on using PAs in new
ways. For example, the two consortia formed in response to the federal Race to the Top
program have released plans proposing to use PAs for a variety of purposes: tracking
student growth, measuring students’ readiness for college and the workforce, and
evaluating teacher effectiveness (PARCC, 2010; SBAC, 2010). Moreover, current plans
propose to move away from traditional paper-based task administration and human scoring
towards online administration and artificial intelligence (AI) scoring. Thus, innovation in the
identification and collection of validity evidence that can support these new and varied
purposes and contexts is required.

Recent initiatives propose to use PAs for both formative and summative purposes.
Formative use occurs when educators utilize assessment results during the instructional
cycle to support continued teaching and learning. For example, PA results might be used to
improve instruction for groups of students, as when evaluating the success of instruction
overall and for different sub-groups; evaluating the success of a newly-implemented
program, initiative, or instructional tool; and planning future instruction. PA results might
also provide information useful for improving individual student learning, including:

• identifying student strengths and weaknesses
• identifying student misconceptions
• designing instructional remediation for individuals or small groups of students
• communicating information about individual student proficiency to students and parents

• identifying and assigning leveled instructional materials

• identifying students for placement into special/tracked classes (including gifted and talented)

Moreover, plans put forth by the two assessment consortia propose to use PAs for summative purposes, or to hold students, teachers, and/or schools accountable for overall student learning once the instructional cycle is complete. In this case, PAs might furnish information about groups of students, as when results aggregated to the classroom or school level are used to evaluate educator effectiveness or support decisions about adequate yearly progress. PA results might also be used to certify individual student readiness for college or career, determine graduation or grade promotion, or assign student grades.

Overlaying the assessment purpose are the stakes associated with that purpose—whether high or low and whether stakes attach to students, educators, or both. Typically, assessments used for formative purposes are associated with lower stakes than those used for summative purposes, although there are exceptions. An example of a high-stakes formative use would be if an assessment were used to identify students for placement into remedial or tracked classes. The consequences of a wrong placement could be dire for an individual student. One could also think about a low-stakes summative use: the PSAT is an assessment administered to individual students (typically in their junior year) to provide information regarding students’ skills in critical reading, math, and science relative to similar skills represented on the SAT and required in college. Although very high performance might qualify students for scholarship opportunities, there are no negative consequences for particularly low performance. Thus, student stakes are relatively low.
Each of these potential testing situations would require support in the form of specific validity evidence relevant to the intended use. In addition, several potential dimensions of score interpretation may be relevant to the validity of a particular use, including (1) whether interpretations are at the level of individual students or groups of students (e.g., classrooms, schools, or districts); (2) whether interpretations are norm- or criterion-referenced; and (3) whether interpretations are about current status or growth. Identifying which of these three dimensions apply to any particular intended use is an important first step in determining appropriate and adequate validity and reliability evidence for supporting a given purpose.

**A Validity Agenda for PA**

Validity is an evolving, evidence-based judgment about the plausibility of a particular test score interpretation. Validation is never complete; rather, validity evidence is accumulated over time and integrated into an overall validity argument (Kane, 2006). Although an investigation of validity should begin with evidence that is relevant to current intended uses and interpretations, assessment use may change over time to accommodate new or unanticipated purposes. When assessments are used in ways originally unanticipated by assessment developers, additional validity evidence relevant to this new use must be collected. Accordingly, additional validity evidence should be collected over time as new information on intended uses becomes available, and as assessments are used in new contexts to support different purposes. We propose a flexible validity agenda that will both organize validity evidence collected over time and move the measurement field forward in terms of best practices in PA. The framework we propose includes consideration of the five sources of validity evidence documented in The Standards for Educational and Psychological Measurement (AERA, APA, NCME, 1999): test content, response processes, internal structure, external relationships, and consequences of use.
Test Content

Validity evidence based on test content refers to any evidence documenting that the target content, processes, and competencies—those identified as necessary to support the intended inferences—are actually present in test items. This means that the content and processes reflected in test items actually align with the knowledge, skills, processes, and abilities they were designed to measure. For PAs, it is also important that these content and process domains be assessed as directly as possible.

To support the validity argument, task development processes (including whether and how classroom teachers provided input into task design) and specified procedures for task administration and scoring should be well documented. In addition, we recommend conducting a formal alignment study to evaluate the extent to which subject matter experts are successful in aligning PA tasks to target content standards and processes (Webb, 2006). To provide evidence that the overall assessment is representative of the domain of interest, we recommend examining the fidelity of the PA to the specifications outlined in the test blueprint. Furthermore, if performance standards identifying graduated levels of performance will be set (e.g., Basic, Proficient, Advanced), documentation of the procedures used to define the performance level descriptors and determine cut scores should also be integrated into content-related validity evidence. Finally, if tasks are designed to align with a specific curriculum, we recommend collecting curricular information, such as instructional modules that the assessments are intended to accompany. For example, if PA tasks are to support interpretations regarding teacher effectiveness, close alignment between the content and processes reflected in the tasks and those targeted by curriculum and instruction is critical. For supporting interpretations about individual students, documentation of students’ opportunity to learn the content and skills is especially important (Messick, 1994). Thus, one could compare the intended curriculum to the enacted
AN EVIDENCE-BASED VALIDITY ARGUMENT FOR PA

curriculum to determine the level of implementation fidelity and identify potential gaps in alignment.

Evidence-centered design (ECD) provides a comprehensive framework for planning, organizing, and documenting content-related validity evidence prior to and during assessment design. Therefore, an ECD approach (e.g., Mislevy, Almond, & Lukas, 2004; Mislevy & Haertel, 2006) may be especially appropriate when planning validity evidence to support the development of performance tasks. In the ECD framework, the layers of domain analysis and domain modeling require assessment developers to establish an assessment argument through clear documentation of the target domain. In ECD’s conceptual assessment framework layer, the assessment argument can then be expressed in structures and specifications for tasks likely to elicit the target knowledge and skills, including administration conditions and scoring procedures (Mislevy, Steinberg, & Almond, 2003). In addition, if the assessment is intended to support criterion-referenced interpretations, the artifacts of the ECD domain analysis process may be used to inform the creation of performance level descriptors and ultimately the standard setting process (Plake, Huff, & Reshetar, 2010). An added benefit of ECD is that careful adherence to design procedures during task development helps to address other aspects of validity in which PAs are typically challenged, such as generalizability of task scores across different tasks, raters, and occasions.

One example of a real-world application of ECD during test development was the identification of ECD “leverage points” within an existing test development process for a large-scale science assessment. In this case, researchers applied an ECD approach for developing rich, contextual scenarios associated with test items and tasks. As part of this process, the research team developed design patterns identifying focal and ancillary knowledge, skills, and abilities as well as fixed and variable task features. Within the design pattern, the team also conceptualized narrative structures, which are similar to storyboards
depicting a problem context that can be associated with multiple items and tasks. These ECD tools allowed test developers to support the development of tests that address difficult-to-assess science topics using rich, highly-contextualized scenarios. Ultimately, engaging in an ECD process prior to task development allowed the research team to document and improve the validity and efficiency of the scenario-based test design, development, and delivery process (Snow et al., 2010).

Validity evidence regarding test content should answer the following types of questions:

1. Are PA tasks aligned with relevant content and process standards?
2. Are scoring rubrics aligned with relevant content and process standards?
3. Are PA tasks aligned with the intended and/or enacted curriculum?
4. Do students have adequate opportunity to learn and practice target content and processes?
5. What construct-irrelevant sources of variability (such as writing skill on a test of historical knowledge) contribute to task performance?
6. Do PA tasks fully represent the construct, as defined by the test blueprint?
7. Do PA task modifications (e.g., extended administration time, student choice of task, allowing students to work in groups) alter score interpretations?
8. If using a vertical scale to capture growth, how valid are the content-based assumptions upon which the scale is based?
9. Does the creation of PA tasks follow a rational approach that promotes content representation and alignment to target standards?
10. Do the content and skill-based expectations represented in the PLDs align with the relevant content and process standards?

**Examinee Response Processes**

Response processes refer to the procedures, strategies, and cognitive behaviors that an examinee engages in while responding to a test item. For PAs, this would include overt or implicit procedures or component skills that are activated by a particular task. For example, a science task designed to measure student ability to design, conduct, and interpret the results of an experiment might prompt the student, either implicitly or overtly, to use the scientific method. Evidence related to students’ thought processes is important because it helps to support the argument that the tasks are eliciting the intended knowledge, processes, and skills. In addition, a close examination of students’ response processes provides insights into the cognitive processes underlying acquisition of the targeted knowledge and skills. Different response processes and performance outcomes reveal different levels of understanding or knowledge and indicate whether students are on the right track or caught up in common misconceptions. Explication of the knowledge-task relationships solidifies the validity argument and is also necessary for crafting rubrics and scoring algorithms.

Such evidence is particularly important for PAs because PA proponents make several claims about the superiority of PAs for measuring certain types of student knowledge and skills relative to more traditional item types, including selected-response items. For example, supporters claim that performance tasks provide more direct and transparent measures of student abilities than selected-response items. Similarly, proponents claim that PAs are able to assess students’ knowledge and skills at deeper levels than traditional assessment approaches and are better suited to measure skills like writing and critical
thinking (Frederiksen, 1984). Such claims constitute validity arguments that must be substantiated.

We recommend collecting information on students’ response processes from cognitive labs or “think alouds,” in which students complete PA tasks while verbalizing their thought processes. This type of study could also examine response processes associated with selected-response items written to the same content and process standards as those on which PA tasks are based to determine which item type best elicits intended processes and skills. If participant pools are sufficiently diverse, examinee response processes could also shed light on possible sources of bias or unfairness. For writing tasks, evidence relevant to response processes might come from planning documents, such as notes, outlines, or first drafts. Within an online testing environment, evidence on examinee response processes might also come from response latency data. For technology-enhanced items, such evidence can be culled from logs documenting an examinee’s interactions with stimulus materials and actions taken (e.g., items clicked on, opened, and manipulated).

Several researchers have conducted cognitive labs in order to determine whether examinees employ strategies and processes intended by the test developers. In one study, for example, examinees were asked to think aloud as they responded to several different types of algebra constructed-response items. Results from that study suggested that students of all ability levels identified relatively few strategies for solving tasks requiring extended responses, and several participants did not even attempt constructed-response items because they could not identify appropriate strategies (Beimers, Burling, & Veazey, 2011). In another study, examinees “thought aloud” as they interacted with technology-enhanced items (TEIs) developed to align with Common Core State Standards in English/Language Arts and Mathematics. Such items were designed to measure skills, such as speaking, listening, and researching, that are difficult to measure using more traditional, static item formats. Results suggested that students were, indeed, applying processes and
strategies intended by test developers. However, results also suggested that unfamiliar functionality associated with these items increased the time needed to respond to TEIs, providing evidence of construct-irrelevant variance (Dolan, Goodman, Strain-Seymour, Adams, & Sethuraman, 2011).

Validity evidence relevant to examinees’ response processes should answer the following types of questions:

1. To what extent do PA tasks elicit the intended processes and skills?
2. What construct-irrelevant sources of variability (such as language load on a test of math) contribute to task performance?
3. Do PA tasks fully represent the construct?
4. Do students from different subgroups respond to PA tasks in a similar manner?
5. Are the defined scoring rubrics consistent with the skills and processes elicited by the tasks?
6. How can PA tasks be created to best discriminate among and provide insight into different thinking or reasoning patterns?

**Internal Test Structure**

Internal test structure refers to the interrelationships, often expressed as correlations or covariances, between performance on different parts of a single test form. For PAs, this would include correlations between scores on two PA tasks of the same type from the same test form (e.g., two or more documents-based history tasks); correlations between multidimensional rubric scores from the same task (e.g. separate organization and mechanics scores on a writing task); or correlations among and between different types of PA tasks within the same test form (e.g., a writing and a speaking task). Evidence regarding
internal test structure is important because it helps to evaluate whether expected interrelationships hold empirically and whether theoretically-based distinctions between different types of PA tasks or different types of rubric scores are meaningful.

We recommend conducting exploratory and/or confirmatory factor analysis and structural equation modeling approaches to investigate interrelationships among different tasks of the same type, among different PA task types, and/or among multidimensional rubric scores from the same task. Such analyses can provide construct-related evidence that tasks exhibit intended and expected structures. Given adequate examinee subgroup sizes, we also recommend conducting multi-group structural equation modeling to explore the invariance of factor structure across certain examinee subgroups. Invariance of factor structure across subgroups would suggest that the tasks are measuring the constructs similarly across groups. Such evidence would help to rule out test or item bias as a possible explanation for potential differences in mean scores across different subgroups.

For example, one study (Lai, Auchter, & Wolfe, 2012) investigated the internal factor structure of a performance assessment of teacher quality that included both teacher portfolios and teacher responses to extended constructed-response items. Portfolio entries were designed to capture pedagogy, whereas constructed response items were intended to measure teachers’ subject-matter knowledge. The researchers tested a theory-based, two-factor model of teacher quality, comprising pedagogy and subject-matter knowledge against a single-factor model of teacher quality. Results from that study confirmed the expected two-factor structure of the assessment, supporting the conclusion that the two assessment portions capture qualitatively different kinds of evidence about teacher quality. Furthermore, slight variations in results across different subject areas and grade levels provided empirical support for the use of separate exams for each subject area and grade-level combination.
Evidence related to internal test structure should answer the following types of questions:

1. Do expected interrelationships among PA task scores materialize?
2. Is the internal PA structure invariant across examinee subgroups?

External Relationships

Another source of validity evidence is the relationship, often expressed as correlations or regression coefficients, between performance on the test and performance on other measures of the same or different constructs. This evidence is called convergent validity evidence when the measures assess the same construct and discriminant validity evidence when the measures assess different constructs. Relationships can be concurrent, as when two measures are administered at the same time, or predictive, as when current performance is related to performance at some point in the future. For PAs, this would include relationships among PA task performance and performance on other assessment types (e.g., traditional selected-response test) measuring the same constructs. Such evidence might also include relationships between scores on PA tasks from two different content or process domains (e.g., reading and math) and relationships between PA task performance and criterion performances that we are interested in predicting (e.g., success in college-level courses).

We recommend conducting correlation and regression analyses to explore relationships between performance on PA tasks and performance on a variety of external measures. For example, a multitrait-multimethod approach (Campbell & Fiske, 1959) could provide both convergent and discriminant validity evidence. Scores on PA tasks can be compared to scores on selected response-type measures of the same construct, and students can respond to PA tasks in multiple domains. Scores on PA tasks should correlate more strongly with other types of measures of the same construct than with scores on other
PA tasks measuring different constructs. Such analyses can help to rule out common method variance (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). In addition, if relevant, we recommend conducting regression analyses to investigate whether scores on PA tasks predict important criterion performances, such as first-year college GPA.

Evidence collected about external test relationships should answer the following types of questions:

1. Do expected external relationships materialize?
2. Does PA task performance predict important criterion performances?
3. What construct-irrelevant sources of variability (such as common method variance) contribute to task performance?
4. Do PA tasks fully represent the construct?

**Consequences of Test Score Use**

Consequences of test score use refer to intended or unintended outcomes that result from using an assessment in a particular way to support specific interpretations. For example, an intended consequence of placement testing is that students are placed in classes where they can receive appropriate instructional interventions. In this case, evidence suggesting that such students perform better than similar-ability students who were not placed in special classes would constitute evidence about intended test consequences. An unintended consequence of placement testing might be that a disproportionate share of minority students are placed into remedial or special education classes. Unintended consequences only affect judgments about validity to the extent that they can be traced back to a source of invalidity in the test—either construct-irrelevant variance or construct under-representation (Messick, 1989). In this case, if the adverse
impact of the test could be traced back to construct-irrelevant variance, this would invalidate the use of the test for placement of minority students.

For PAs, consequences of test score use might include changes in teachers’ instructional practices and in students’ learning. For example, proponents of PAs claim that their adoption can improve teaching and increase student engagement and motivation (Hancock, 2007), and that involving teachers in writing and scoring PAs can serve as a valuable professional development experience (Borko et al., 1997). Assessments designed for formative use often claim to provide information that teachers can use to improve learning and increase student achievement. Such claims constitute validity arguments that must be substantiated.

The specific validity evidence to be collected will depend on proposed test uses and interpretations. For example, if PAs are intended to improve the quality of instruction, we recommend investigating teachers’ instructional practices, perhaps by administering teacher surveys, conducting interviews and focus groups with teachers, and/or conducting classroom observations. If the introduction of PAs is intended to serve as a form of professional development for teachers, we recommend surveying and interviewing teachers about their professional development experiences. If PAs are designed to stimulate greater student engagement and motivation, we recommend surveying or interviewing students about their task-related motivation and conducting classroom observations to monitor and document levels of student engagement while performing tasks. Finally, if PAs are intended to serve formative purposes (i.e., to provide information useful for improving learning), we recommend collecting evidence demonstrating that, when properly used, information provided by PAs contributes to improved student learning. Evidence pertinent to the consequences of test use should be collected over the life of the testing program, particularly as tests are used in new ways. However, any consequential evidence collected ought to be evaluated relative to whether negative consequences can be traced back to a
source of invalidity in the test—either construct-irrelevant variance (e.g., insufficient testing time) or construct under-representation.

Validity evidence regarding consequences of test score use should answer the following types of questions:

1. Do PA tasks achieve intended consequences for teaching and learning?
2. Do PA tasks result in unintended consequences for teaching and learning?
3. Can unintended, negative consequences of test score use be traced back to sources of invalidity in the test?

**Scoring**

Scoring of PAs raises validity issues that cut across the five sources of validity evidence identified by the Standards. The impact of scoring on validity is typically discussed in the context of reliability—in particular, the reliability of PA scores as evidenced by levels of consistency both across and within raters (Dunbar et al., 1991). However, one could also evaluate the quality (i.e., clarity, specificity) and relevance of the rubrics being applied, the accuracy of rater-assigned scores (e.g., the extent of agreement with expert ratings), or the fidelity of rater cognition to a particular model of scorer behavior using cognitive labs (see, for example, Wolfe, 1997; Wolfe & Feltovich, 1994; Wolfe, Kao, & Ranney, 1998). Thus, several scoring issues are clearly related to validity.

The topic of automated scoring makes the connection between scoring and validity even more apparent. Automated scoring is a key technology supporting the increased use of performance assessments for both formative and summative purposes. Automated scoring has the benefits of being able to significantly reduce score turnaround time and cost, and increase score consistency. Previous research studies on automated scoring systems report high agreement rates between automated scores and scores assigned by human raters.
(Attali, 2004; Burstein & Chodorow, 1999; Landauer, Laham, & Foltz, 2003; Landauer, Laham, Rehder, & Schreiner, 1997; Nichols, 2004; Page, 2003). Although the capabilities of automated scoring are widely recognized, potential validity issues associated with the use of automated scoring for PA tasks need to be carefully considered.

Current automated scoring approaches, which differ from one another in terms of the underlying scoring algorithms and techniques, are all designed to predict human scores. However, since an automated scoring system can not read or comprehend responses in the same way a human does, it is unable to experience the decision-making process used by a human rater when evaluating a student response. Instead, an automated scoring system attempts to mimic human cognition by identifying and extracting a number of response features that correspond to different levels of quality, and aggregating those features into scores through the application of a scoring model. The choice of features and applied aggregation rules are important evidence which speaks directly to the validity of automated scores. Decisions regarding what the features are and how they should be aggregated for a particular performance task should be rooted in a deep understanding of the construct being measured by the task. Validity evidence can come from asking content experts to define the qualitative features, confirm the aggregation rules, and verify the connections between automated scores and observed performances.

Scoring engines are trained and calibrated with the use of human scored responses, often collected in a pilot or field test. The reliability and accuracy of human scores assigned to papers used to train and calibrate the engine are therefore a crucial contributor to the quality of automated scores. For this reason, evidence on the accuracy of human raters (i.e. agreement with expert raters) and the representativeness of the training paper sample is also needed to evaluate the meaningfulness of automated scores. Information regarding how human raters are trained and how scoring quality is managed should also be documented. In addition, questions including whether human raters follow the provided
rubrics faithfully during real-time scoring, what features they actually use when they evaluate the responses, and how they arrive at summary scores can be answered by inviting human raters to cognitive labs or think-aloud studies.

Automated scoring typically presumes an on-line testing platform. As a matter of fact, an automated scoring system should be considered as part of a computer-based testing system in which all the components are intricately connected to one another (Bennett & Bejar, 1998). The interplay among the different components of the assessment system defines the validity of the resulting scores. As Bennett (2011) states, “the dependencies among components are particularly salient with respect to the connection between the computer interface and automated scoring” (p. 11). A computer interface sets up the task and imposes constraints that restrict the ways in which students interact with computers and enter their responses. Additional interface constraints tend to increase the predictability of students’ responses and, in turn, enhance the accuracy of automated scores. On the other hand, additional interface constraints may limit the opportunities for students to demonstrate extended performances and, in the end, change the nature of the task. Therefore, a well-designed interface should narrow the open-endedness of admissible responses while still getting at the problem-solving nature of performance tasks.

Validity evidence regarding automated scoring should answer the following types of questions:

1. What are the response features measured by the automated scoring system? Are they aligned with the construct of interest? How are they identified and extracted?

2. How are the features aggregated into scores? Are the aggregation rules aligned with the construct of interest?
3. How reliable and accurate are the scores assigned by humans to the anchor papers used to train the engine? How representative is the training sample in terms of the range of possible score points?

4. What is the relationship between automated scores and scores of expert raters assigned under ideal conditions? Is the relationship invariant across different groups?

5. What is the extent of agreement between automated scores and scores of human raters assigned under operational scoring conditions? How do automated scores and operational human scores correspond to an external criterion such as students’ scores on a selected-response test?

6. Is the automated scoring approach subject to manipulation or sensitive to creativity? Does the scoring approach have any negative impact on learning and instruction?

7. Does the computer interface introduce extraneous factors that evoke types of performances that are not intended by the task developers?

**A Reliability Agenda for PA**

Reliability is a necessary, though insufficient, condition for validity. In other words, a test cannot be valid if it does not produce scores that are consistent and relatively free from error. Thus, an investigation of validity should include an examination of reliability. The strongest criticism of PAs is that they suffer lower reliability and generalizability than selected-response items, primarily because of differences in efficiency between the two task types. In particular, one hour of testing time buys you far fewer performance tasks than selected-response items. Thus, in exchange for greater depth of content coverage, PAs compromise breadth of content sampling (Messick, 1994). Generalizability studies of PAs
have found that significant proportions of measurement error are attributable to task sampling, manifested in both person-by-task interactions and person-by-task-by-occasion interactions in designs that explicitly model the occasion facet (Shavelson et al., 1999).

Extended performance tasks are also more difficult to standardize than are selected-response items (Haertel & Linn, 1996). Lack of standardization characterizes task construction, task administration, and scoring conditions. When implemented in the classroom, PAs are susceptible to being modified in a number of ways, including allowing extended administration times, providing students with choices about which tasks they will respond to, and allowing students to complete the tasks in groups (Haertel & Linn, 1996). Human scoring introduces a certain amount of error, which constitutes another source of unreliability. When human scorers are practising teachers, who typically do not undergo the type of rigorous training and monitoring that expert scorers complete, one can expect the magnitude of scoring errors to be even larger.

Generalizability theory (Brennan, 2001) provides a way of investigating measurement error that is particularly appealing for performance tasks. Generalizability theory permits one to decompose measurement error into its constituent sources and quantify the proportion of error attributable to each facet of the measurement procedure. In other words, generalizability theory allows one to determine the relative impact of various aspects of PA tasks on the consistency of student scores. In the case of PAs, such aspects may include tasks, forms, testing occasions, raters, rubrics, and all possible interactions among these facets. Generalizability theory can also accommodate complex measurement situations, such as when scores are used to support group-level interpretations and when measuring multidimensional constructs.

In addition to estimated variance components and reliability indices provided by generalizability and decision studies, other indices of task reliability can provide
supplemental information. Such indices might include rater consistency and accuracy, Cronbach’s alpha, conditional standard errors of measurement, and the accuracy and consistency of any classification decisions based on scores.

Evidence related to the reliability of PAs should answer the following types of questions:

1. What proportion of variance in scores is attributable to each of the various facets of the measurement procedure? (e.g., persons, tasks, forms, occasions, raters, rubrics, etc.)

2. How reliable are PAs? In the case of multidimensional measurement, how reliable are individual components and composite scores?

3. How many more tasks, raters, occasions, etc. would be required to achieve a minimally acceptable level of reliability?

4. How consistent and accurate are rater scores?

5. What is the magnitude of measurement error at various points along the score scale?

6. If growth interpretations are desired, how reliable are growth scores?

7. If relevant, how consistent and accurate are classification decisions?

8. If interpretations are at the group-level, how reliable and consistent are scores and decisions about groups of students?

**Integrating Validity Evidence**

As Kane (2006) observed, “the validity argument is to provide an overall evaluation of the intended interpretations and uses of test scores by generating a coherent analysis of
all of the evidence for and against the proposed interpretation/use, and to the extent possible, the evidence relevant to plausible alternate interpretations and decision procedures” (summarizing Cronbach, p. 22). Thus, any one piece of evidence cannot be used in isolation in supporting a given use. Rather, the network of inferences connecting interpretations with relevant confirmatory or disconfirmatory evidence should be outlined in its entirety so as to facilitate judgments about the plausibility of those claims.

According to Kane (2006), this process involves two distinct steps. First, one must build the interpretive argument for the assessment. The interpretive argument specifies the relationship between observed performances and the interpretations one wants to support. As Kane argues, these relationships can be expressed as a series of “if-then” statements. For example, “if the examinee’s lab report demonstrates proper use of lab equipment and scientific testing procedures, then the student should earn a score of 4 or higher.” “If the student earns at least a 5, then he or she should be considered proficient in high school biology.” Such statements begin to make apparent a number of assumptions underlying each proposed interpretation. The validity argument, then, should evaluate the plausibility of each of these assumptions by integrating evidence collected from a variety of sources that are relevant to each interpretation.

For example, to investigate the first assumption above, one might combine expert evaluation of appropriateness of rubrics used to score the lab report (content-related validity evidence) with estimates of human scoring consistency (reliability evidence) and evidence related to fidelity of scorer cognition during the scoring process to a specific model of rater behavior (i.e., a form of scorer accuracy). To construct the validity argument, one would make an overall judgment about the plausibility of the scoring inference based on evidence related to the adequacy of the rubric and the consistency and accuracy of scoring processes. As Kane (2006) notes, inferences related to generalization from observed scores
to universe scores, extrapolation from universe scores to the level of the skill, and decisions based on the inferred level of skill could be similarly sketched out and evaluated.

**Summary and Areas for Future Research**

Given recent changes in the way that PAs are used and the types of inferences they need to support, it is imperative to take a systematic approach to conceptualizing, planning for, and collecting validity and reliability evidence. A validity and reliability research agenda, such as that described above, offers a way to organize diverse sources of validity and reliability evidence accumulated over time. Such an agenda identifies validity and reliability issues unique to PAs, as well as potential research questions and possible sources of evidence.

Several outstanding research questions should be answered in the coming years as empirical evidence connected to new PA systems begins to be collected: How reliable are the types of PAs that can be administered efficiently on a large-scale? How accurately can automated scoring systems handle complex student responses? To what extent can performance on one or two tasks be treated as representative of and generalizable to performance in the domain?

It remains to be seen whether the assumptions underpinning proposed new interpretations will be borne out. For example, in order to use PA scores to support inferences about teacher effectiveness, a complex chain of assumptions must be substantiated, beginning with inferences about the meaning of individual student scores and ending with inferences about teaching skill that are based on aggregations of those scores. Clearly, if every link in this chain must be separately evaluated, a substantial body of evidence collected from diverse sources will be necessary. Outlining a potential research agenda capable of building a validity argument is the first step in what is sure to be a long journey.
References


