Validation of Performance-Based Assessments

M. David Miller, University of Florida

Robert L. Linn, University of Colorado

Using Messick's (1995, 1996) framework for validity, six aspects of construct validation are outlined to guide the validation of performance-based assessments: content, substantive, structural, generalizability, external, and consequential. Each aspect is discussed, with the focus on studies that could be conducted within the context of a large-scale educational assessment. Also discussed are the issues that affect construct validation within that context, and recommendations for future areas of study are outlined. *Index terms: accountability, generalizability, performance-based assessment, statewide testing, validity.*

Aschbaker (1991) pointed to an increase in the use of performance-based assessments (PAs) in educational assessment at the state level. Although the trend toward increased state-mandated PAs has slowed, the use of PAs—often in conjunction with more traditional multiple-choice (MC) assessments—continues to be an important part of many state assessment programs. PAs are also beginning to be used in credentialing examinations, such as the use of classroom observation in the certification of teachers in Florida.

There are many reasons for the use of PAs. One is a concern about the possible unintended negative effects of MC assessments (Shepard, 1989), such as the belief that they lead to narrower curriculum and teaching to the test. PAs are believed to be more consistent than MC examinations with a reconceptualization of teaching and learning as a richer, context-bound experience that does not rely solely on rote skills (Perkins & Salomon, 1989).

An advantage of PA is that it is done with methods that more accurately reflect the teaching and learning process, rather than with "a summative measure of the effects of schooling" (Wiggins, 1989, p. 41). Wiggins suggested that educational reform could best be accomplished through changes in assessment programs, because they "determine what teachers actually teach and what students actually learn" (p. 41). High-stakes assessments are often implemented with an assumption that teaching to the test will be one of the positive effects of the assessment program.

The issue of the effects of assessment on teaching and learning has only recently been examined. Validity evidence of PAs in a high-stakes environment is continuing to accumulate, informing the uses and inferences that can be made from PAs.

Validity of PAs

The Standards for Educational and Psychological Testing [American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education, 1999] defined validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Although few would dispute this definition or the importance of considering validity as a unified concept (Messick, 1989), actual

Applied Psychological Measurement, Vol. 24 No. 4, December 2000, 367–378

criteria for examining validity vary widely. AERA et al. discussed five sources of validity (based in part on Messick, 1989): evidence based on (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of testing.

Messick (1995, 1996) argued that validity is a unitary concept. However, Messick indicated that the validation of PAs, like other forms of assessment, should incorporate six aspects of construct validity: content, substantive, structural, generalizability, external, and consequential. The content aspect of validity focuses on the relevance and representativeness of the assessment's content. For the substantive aspect, the focus is on the processes used by examinees when they respond and on the consistency of those processes with the construct the assessment is designed to measure. Hence, evidence is sought that demonstrates that tasks lead examinees to engage in the intended cognitive processes, and that the influence of construct-irrelevant factors is minimized. The structural aspect addresses the adequacy and appropriateness of scoring and scaling. Generalizability focuses on the replicability of results across multiple levels of facets of the assessment procedure (e.g., raters and tasks). In the external aspect, the convergent and discriminant evidence is examined that shows the relationship between the assessment and other measures and constructs. The consequential aspect examines the degree to which assessments have both the intended positive effects and plausible unintended negative effects.

Messick (1989, 1995, 1996) provided a broad framework for examining validity. However, Shepard (1993) argued that the breadth of the conceptual framework tends to overwhelm the practitioner in many situations. Thus, it is important to prioritize validity concerns, as well as to clarify distinctions between different aspects of validity. For example, a change in content of an assessment can affect scoring methods, generalizability of the assessment, or consequences of its use. Shepard discussed ways of prioritizing validity questions based on the purpose of the assessment.

Six Aspects of Construct Validity

Content

According to AERA et al. (1999), evidence based on content is "obtained from an analysis of the relationship between a test's content and the construct it is intended to measure" (p. 11). Validation studies that examine evidence of the content are typically done through an analysis of the content of the tasks, the curriculum, and the domain theory (Messick, 1996). Typically, this part of construct validation relies on expert judgments about the boundaries of the construct, the curriculum, and the skills and content measured by the tasks (Crocker, 1997). However, it can be expanded to include correlational analyses of the relationships of assessments based on content (Stecher et al., 2000).

Efforts in constructing PAs for state assessment programs typically are based on state-adopted content standards that specify what it is that students should learn and teachers should teach, as well as what should be assessed. Content standards, though neither curriculum nor fully specified sets of specifications (blueprints) for an assessment, are expected to provide the basis for both curriculum and assessment specifications. Much has been written about the need for alignment among content standards, curriculum, instruction, and assessments (e.g., Linn & Herman, 1997; McLaughlin & Shepard, 1995). However, a specific meaning of "alignment" and how it should be determined are seldom discussed. The concern here is issues of assessment alignment and the constructs that are invoked by content standards. In Messick's (1989) framework, questions of alignment are fundamentally about the content aspect of validity.

Content analysis can occur in two phases: (1) initial task/assessment development, and (2) review after the assessment has been constructed. During Phase 1, a blueprint provides evidence of

the content and skills being assessed. In standards-based assessments, the blueprint is derived by an elaboration and more-detailed specification of the types of tasks/activities that are consistent with the intent of the content standards. The blueprints should clarify the constructs that are intended and the relationship between the construct and individual tasks. Relative weightings of content and skills within the blueprint are important for showing the relevance or criticality of specified areas; they are crucial to defining the inferences that can be made from the assessment. For example, two mathematics assessments covering the same content area can lead to different inferences, depending on the relative weighting of skills and content within the broader construct of mathematics.

Phase 2 provides evidence regarding the judgments of the assessment developers. Unless it is carefully done, the development process can become more focused on characteristics of particular tasks (e.g., item analysis, review of individual items), rather than on the breadth of the total assessment. Review by individuals not involved in the development process focuses beyond the characteristics of particular tasks and onto the construct, to examine the breadth of the assessment content and its relationship to the construct.

Examining the overall breadth of the assessment becomes especially critical when the assessment has few items and/or other threats to validity (e.g., item homogeneity), as these can lead to a construct interpretation that is too narrow. For example, a writing assessment consisting of a single prompt might be judged as inadequate in the content aspect because different types of writing assessments are not strongly correlated (e.g., expository and narrative). Thus, when the assessment is taken, a good measure of "writing" is not obtained. Questions also might be raised about the representativeness and adequacy of a single prompt for any subdomain of the content. Content reviews allow definition of the boundaries of the inferences that can be made from assessment scores.

Substantive

This aspect examines substantive theories and process models (cognitive processing) that are used as a basis for performance on tasks. According to AERA et al. (1999), "theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (p. 12). As tasks need to be representative and relevant to content and skill specifications, so processes used in completing the tasks need to be representative and relevant to the processes that constitute the construct of interest. Content standards adopted by states generally also include processes, as shown by words such as "analyze," "describe," "evaluate," "explain," "demonstrate," "collect," "organize," "solve." The validity of a PA for assessing the constructs implied by these processes needs to be evaluated, as do the content aspects of the constructs.

As in the case of content, expert judgments about the process models and their relationships to the construct are needed. In addition to reviewing the tasks, the review of the scoring rubrics (i.e., scoring guidelines for making subjective judgments) needs to show that scores are based on the successful completion of a process.

In addition to expert judgment, Messick (1996) noted the need "to accrue empirical evidence that the ostensibly sampled processes are actually engaged by respondents in task performance" (p. 9). Process model studies could include an analysis of examinee responses using "think alouds" (a qualitative method of judging the process used to solve a problem by asking students to solve a problem orally), or empirical investigations of process models. The proposed Voluntary National Test included "cognitive labs," which illustrated the use of think-aloud interviews as part of the initial item development process (e.g., Wise, Hauser, Mitchell, & Feuer, 1998). Empirical investigations of process models could include studies examining the relationship of performances across different

process models or experimental studies that teach how to do the relevant process in a pretest-posttest design.

Structural

The structural aspect of construct validity examines the scoring system as it relates to the construct domain. According to AERA et al. (1999), "analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (p. 13). With PAs, multiple options exist for scoring complex, open-ended responses. Scoring often is done with rubrics that focus on pieces of the assessment (e.g., analytic or primary trait) or the performance as a whole (e.g., holistic). Unlike traditional paper-and-pencil assessments, the selection of a scoring method is central to understanding the inferences that can be made from PAs (Miller & Legg, 1993). As with content and cognitive processing models, a scoring method helps define the boundaries of the construct being measured and should be subject to expert review.

The scoring method needs to be consistent with the construct domain. Additionally, the implementation of rigorous and systematic scoring procedures is crucial to obtaining comparable scores across scorers (Haertel & Linn, 1996). However, strong models exist for reliable scoring, which include multiple readers, anchors or benchmarks, adjudication, training, and calibration checks for drift—making scoring almost routine within the context of large-scale assessment. These models have led to high levels of rater consistency (Brennan, 1996; Miller, 1998; Shavelson, Baxter, & Gao, 1993).

The appropriateness and adequacy of scaling and equating procedures are also relevant to the structural aspect of validity. When constructed-response and MC items are scaled together using item response theory, for example, an evaluation of the adequacy of that scaling is needed, as is the degree to which the relative weights for the two types of items are consistent with the construct interpretation of the results.

Generalizability

In this aspect, the replicability or consistency of assessment results across multiple levels of random facets of an assessment is examined, in order to understand the boundaries of the construct (Kane, 1982). Traditional MC assessments might only have the random facet of items; consistency can be measured with Cronbach's α or other internal consistency indices. The number of random facets creating error is typically larger for PAs. Perhaps the most frequently occurring designs for examining generalizability involve raters and tasks as sources of error. However, other facets can be included, such as occasion or students (in the case of aggregate results for schools).

Generalizability theory is the generalization of Cronbach's α for designs that are more complex than a fully crossed persons-by-items design. It is based on the application of analysis of variance models and random variance components to estimate universe score variance, and relative and absolute error variance to examine the consistency of the assessment procedures under different universes of generalization (i.e., different conditions of raters and tasks). [For a more thorough explanation of generalizability theory, see Brennan (1992, 2000); Cronbach, Gleser, Nanda, & Rajaratnam (1972); Shavelson & Webb (1991).] PA generalizability has been a focal point for the examination of validity (Linn, Baker & Dunbar, 1991; Messick, 1995, 1996) as well as reliability (Brennan, 1996; Miller, 1998; Shavelson et al., 1993), especially because assessment conditions are usually more complex than is allowed by traditional measures of consistency.

Results of generalizability studies are mixed concerning raters using scoring rubrics. However, as Brennan (1996) noted, mixed results can be attributed to poor results from older studies; current

scoring methods have been refined enough so that "often methods can be found to increase rater reliability" (p. 42). Miller (1998) examined variance components, error variances, generalizability coefficients, and dependability indices across 40 state-mandated assessment programs in four states. With few exceptions, rater variance components and their interactions were relatively small compared to the examinee-by-task interaction, resulting in fewer raters needed (generally 1 or 2) to achieve acceptable values of error variances, generalizability coefficients, and dependability indices.

In contrast, variance components associated with person-task interaction can be large, requiring multiple tasks for adequate levels of generalizability (relative) or dependability (absolute). For person-by-task interaction, results are consistent, suggesting that "methods do not exist for increasing task reliability" (Brennan, 1996, p. 42). Even when tasks are carefully constructed and field tested, examinees respond differently to different items, resulting in a high variance component.

However, several studies have found that different numbers of tasks are needed to ensure adequate levels of generalizability. Miller (1998) found that the number of tasks required to achieve generalizability coefficients of .70 or higher ranged from 2 to 10 in the same study, which was lower than the number of tasks reported necessary by Shavelson et al. (1993; results were reported for a generalizability of .80, but even at the same level of generalizability more tasks were needed). Future studies should focus on the characteristics of tasks/assessments affecting the magnitude of the person-by-task interaction variance component.

The usual explanation for a large person-by-task variance component is task (or item) heterogeneity (Shavelson et al., 1993). Only two solutions can reduce the error associated with task heterogeneity: (1) the number of tasks can be increased (as is typically done with MC assessments), and (2) the construct and tasks can be defined more narrowly. The first solution is impractical because of assessment time limitations, combined with longer times needed for completing performance-based tasks. In the second solution, constructs would have to be more context-bound, and consequent inferences that could be made from the tasks would be narrower. Depending on the assessment purpose and the breadth of the inferences that need to be made, this might not be a good solution. That is, it might not be reasonable to provide scores on "narrative writing" as a separate construct from other types of writing.

Other explanations of larger variance components are also possible. For example, both Miller (1998) and Brennan (1996) reported fewer tasks being needed to achieve comparable levels of generalizability than did Shavelson et al. (1993). Miller's and Brennan's studies were based on statewide high-stakes assessments, whereas Shavelson et al.'s was based on a locally administered science assessment. High stakes could lead to instructional alignment or other classroom practices that reduce the person-by-task variance component. An alternative explanation is that the refinements in the development process for a large-scale operational assessment program reduced the magnitude of person-by-task interaction.

Miller (1998) reported that the number of tasks required for comparable levels of generalizability varied by the type of task. Longer, more complex performance-based tasks required fewer tasks for adequate levels of generalizability than did open-ended, short responses. In Connecticut, as few as two tasks were required with complex, extended tasks for the literature and interdisciplinary assessments. In contrast, Alabama's algebra and geometry assessments, with shorter, open-ended tasks, required five to ten tasks to reach comparable levels of generalizability (Miller, 1999). Thus, there is a need for better understanding of the sources of person-by-task interaction and how it can be reduced as an error source for PAs.

Technical concerns, such as the estimation of generalizability with incomplete or missing data, also should be addressed. To the extent that analyses are based on extant data, estimation of

generalizability is problematic with missing data or incomplete unbalanced designs. In a simulation study, Harrison (1998) found that missing data led to substantial underestimation of internal consistency with a person-by-task design, except under the rare condition that data were truly missing at random when using listwise deletion. However, alternative estimation procedures, which can be routinely operationalized within the framework of generalizability theory, led to unbiased estimates. Harrison (1998) found that strategies based on the estimation of variance components from available data provided reasonable accuracy and precision in all situations. This research should be extended to examine designs with more than one facet of error, as well as other conditions of missing data for the person-by-task design.

Although evidence is beginning to accumulate on the generalizability of PAs across grade level and content area (Brennan, 1996; Miller, 1998), there is little evidence about generalizability at the group level. Because accountability is often at the school level, generalizability studies are important for understanding the use of PAs as an accountability measure. Cronbach, Linn, Brennan, & Haertel (1997) discussed generalizability issues for aggregates that pose new challenges. The designs that are most efficient at the school level might not be those that have the most desirable properties at the individual student level, and vice versa.

Group means are generally assumed to be more reliable than individual scores; however, Brennan (1995, 1996) pointed out that this is not necessarily true. The universe score variance is smaller for groups than for individuals, because the equivalent formula for students contains an extra term (schools plus student within schools, versus schools). In addition, the error term associated with students contains fewer sources of error; students within schools and their associated interactions are a source of error for examining schools, which results in a smaller error variance for students. On the other hand, several variance components in the school error estimate are divided by the number of students within schools, contributing to a smaller error variance for schools. As a result, not only is the universe score variance smaller for schools, but the error variance could be larger or smaller, depending on the magnitude of the students' variance component and the number of students in the study.

Miller (1998) investigated the generalizability of school means in 23 assessment programs across four states. Across the programs, the error variance associated with schools was relatively small (compared to students), with as few as 50 students per school, but was not consistently smaller when there were 20 students per school. The magnitude of the universe score variance varied substantially across assessment programs. Sometimes the magnitude of the universe score variance was so small that any comparison of schools was rendered meaningless. However, the standard errors were still small enough to interpret school means in relation to some criterion when there were enough students. More studies are needed to clarify and extend the results of this study. Empirical studies investigating issues in the generalizability of school-level indices derived from student performance assessments are also reported by Burton (1998), Candell & Ercikan (1994), and Yen (1997).

External

According to AERA et al. (1999), "analyses of the relationship of test scores to variables external to the test provide another important source of validity evidence" (p. 13). This part of validation is used to test whether assessment relationships are consistent with the knowledge base and theory for the construct. Thus, a pattern of high, moderate, and low empirical relationships is examined, based on theoretical expectations or hypotheses. "That is, the constructs represented in the assessment should rationally account for the external pattern of correlations" (Messick, 1996, p. 11). Further

evidence can be drawn from a more fully elaborated theory, providing a stronger basis for score interpretation.

Convergent and discriminant evidence originated from Campbell & Fiske (1959), with their multitrait-multimethod matrix. Convergent evidence is based on correlations that should be high due to a shared construct; discriminant evidence is based on correlations that should be low due to different constructs. Messick (1996) extended this definition: "convergent evidence signifies that the measure in question is coherently related to other measures of the same construct as well as to other variables that it should relate to on theoretical grounds. Discriminant evidence signifies that the measure is not unduly related to exemplars of other distinct constructs" (p. 12).

Often, PAs are not based on a well-defined theory. However, several hypotheses about PAs are consistent with the literature. First, performance should be related to instructional effects. Much of the impetus for PAs is that they should mirror the teaching and learning process and provide a better measure of accountability (Wiggins, 1989). In addition, all achievement measures should be sensitive to instructional effects. Second, construct variance should be higher than method variance (Campbell & Fiske, 1959). This is important, because there is usually a range of methods that can be used to measure the same construct. Third, PAs should be fair, not giving one subpopulation an advantage over another based on construct-irrelevant factors (Bond, 1995; Linn, Baker, & Dunbar, 1991; Miller & Legg, 1993).

Limited data are available on each of these hypotheses. States reported increases in achievement across time on PAs (Blank & Engler, 1992). Of course, this type of data does not examine whether students do better on particular tasks depending on the form or content of the assessment or whether learning really occurs on the broader construct (Miller & Seraphine, 1993). In addition, teachers reported aligning their curriculum with PAs.

A few studies have addressed different forms of assessment. For example, the science State Collaborative on Assessment and Student Standards, sponsored by the Council of the Chief State School Officers (CCSSO), developed assessments based on a common unit while using multiple methods of assessment (Blank, 1989). Table 1 shows the correlations from one form at the high-school level of a large-scale field test. This assessment included MC items from a common core across science content areas [Core (MC)]. The remaining instruments were unique to one content theme or unit. The assessment included 15 MC items [Form (MC)], two written responses to questions (one short answer and one extended or long answer), and descriptions of two hands-on events (Tasks 1 and 2).

Table 1
Correlations of Form 506 Scores From Phase 2 of the CCSSO Science Assessment

Score	Core (MC)	Form (MC)	Short Ans	Long Ans	Task 1
Form (MC)	.35	—			
Short Ans	−.12	−.11	—		
Long Ans	−.12	−.07	.47	—	
Task 1	−.01	.01	−.04	.05	—
Task 2	−.10	.09	−.01	.06	.76

The correlations in Table 1 suggest that the method of measurement had a strong effect, because correlations were of moderate size only when the assessments used the same method and were low when there was no common method. On the other hand, cognitive level and content differences were not specifically examined for this dataset. The correlation of $r = .35$ between the two MC

tests was moderate, as was $r = .47$ between the long- and short-answer items, and there was a high correlation ($r = .76$) between Tasks 1 and 2. All other correlations were essentially zero. These findings were consistent across multiple forms at the elementary, middle-school, and high-school levels. Thus, the method variance was stronger than the content theme variance [the first MC test—Core (MC)—was not part of the unit]. Methods of assessment should be examined further to determine whether constructs should be more context bound, based on method of measurement.

Consequential

The focus of the consequential aspect of construct validity is the intended and unintended consequences of test use and their impact on score interpretation and use. This aspect has been debated over the last few years. Shepard (1997) noted that "the importance of attending to testing effects as a part of validity is not a new invention" (p. 6). It can be traced to Cureton (1951) and Cronbach (1971) under different terminologies. Shepard (1993, 1997) argued that consequences are an integral part of validity because they affect the overall construct use and interpretation of test scores.

In contrast, Popham (1997) argued that the definition of validity in AERA et al. (1985) was adequate, and that the addition of consequences to that definition "will lead to confusion, not clarity" (p. 9). However, Popham also argued that "consequences should be systematically addressed by those who develop and utilize tests, but not as an aspect of validity" (p. 9). Thus, the debate (see also Linn, 1997; Mehrens, 1997) centers around the definition of validity, although it is agreed that examining consequences of test use is important.

AERA et al. (1999) adopted consequences as one source of evidence for validity. A distinction needs to be drawn between consequences of assessments that do not affect the inferences and uses, and consequences of the assessments that do affect the inferences or uses. The former is not a part of validity, whereas the latter is. For example, a school district might invest in new instructional and curricular materials that would address the construct measured on a PA. However, the consequence of the district's expenditure is not a part of validity until the materials are used by teachers to affect student achievement on the PA in such a way that it alters the interpretations or uses of the assessment. Thus, consequences should be linked to actual changes in interpretations of scores or uses of the assessment to address the issue of validity. That is, consequences cannot simply follow the assessment. They should be prior to the interpretation or use of the assessment as a part of the validity framework. For state PAs, consequences during the current year usually have effects on the assessment interpretations both during that year and later.

Consequences can be intended or unintended. Intended consequences might include changes in the instructional and curricular practices of teachers that lead to better learning environments for students (Linn & Baker, 1996). Unintended consequences might include bias in the assessment, leading to misinterpretations for some subpopulation (Bond, 1995).

Few empirical studies have been conducted that examine the consequential aspect of validity. Miller (1999) conducted a study with CCSSO, examining the perceptions of teachers about the consequences of state-mandated PAs in three areas: (1) attitudes and instructional practices toward state PAs, (2) attitudes and instructional practices toward classroom and other PAs, and (3) professional development in the area of PAs.

Table 2 shows teacher-reported attitudes and practices toward state-mandated PAs in five states. Attitudes and self-reported practices were positive, as can be seen from the means; the midpoint of the scale is 2.0. In almost all assessment programs, teachers agreed that they had (1) taken steps to align their instruction with the state PA, (2) supported their state's efforts to improve instruction through their implementation of mandatory assessments, and (3) received support and

encouragement from the school administrators. With the exception of the geometry program in Alabama, teachers also agreed that state PAs had a positive impact by leading teachers toward a common direction with respect to curriculum. Connecticut, Delaware, and North Carolina also reported enhanced professional growth and development and a positive effect on learning.

Table 2
Teacher-Reported Attitudes and Practices Toward State-Mandated Assessment (Percent Responding *Agree* or *Strongly Agree*) From Alabama (AL1 = Algebra, AL2 = Geometry), Connecticut (CT), Delaware (DE), North Carolina (NC), and Rhode Island (RI)

	AL1	AL2	CT	DE	NC	RI
Teachers who report they...						
Align instruction with assessments	75.9	78.8	77.2	83.2	92.7	54.5
Support state's efforts in assessment	68.5	39.4	65.2	73.9	66.0	49.6
Believe administrators support state efforts	67.8	48.2	84.0	84.3	88.5	—
Teachers who report that state-mandated assessments...						
Have positive effect on learning	33.8	11.3	47.4	52.4	54.7	—
Increase student confidence	27.7	9.1	34.4	37.6	37.2	—
Accurately reflect student performance	27.3	13.1	28.7	28.7	22.9	—
Narrow classroom curriculum	28.7	45.2	42.8	36.2	46.8	—
Enhance professional growth	35.0	24.8	44.2	48.9	38.5	—
Provide a common direction for curriculum	58.5	42.2	47.4	57.6	50.5	—

Note. "—" indicates that the state did not use the question.

Although the majority of teachers in assessment programs had positive attitudes and practices toward state PAs, several other trends should be noted. First, the state-mandated assessments were not considered to be an accurate reflection of student performance, and the tests were not judged to increase student confidence. However, neither of these findings should be surprising, given the purposes of the assessments and the time needed to complete them. The assessments were usually intended to give supplemental information and, consequently, they did not reflect everything that students learned. Only a small view of student performance was provided, given the limited time for administration of the assessments. The second trend was lower ratings in the Alabama geometry assessment program. These lower ratings could have been due in part to low scores on the tests. Another factor that might have affected teacher ratings was that Algebra I and Geometry were the only high-school courses that had state-level accountability testing attached. Clearly, more information (such as focus groups; Chudowsky & Behuniak, 1998) should be collected to attempt to understand this phenomenon.

State-mandated PAs have many consequences. Miller (1999) found that Rhode Island, the only state in its first year of implementing state PAs, had teachers spending significantly less time engaging in classroom PAs. Another effect on classroom instruction was the widespread use of teacher-developed PAs. This finding might also have an intended negative side effect, in that a relatively low percentage of teachers reported professional development in the last year, suggesting

that more professional development is needed—especially in the areas of assessment development and evaluation.

On the other hand, it should be apparent that Miller's (1999) study examined only a small part of the consequences of state-mandated PAs. This study focused only on intended consequences and uses. Moreover, the evidence was limited to teacher self-reports. The surveys were used to examine teacher perceptions of consequences on professional development, instructional practices, and attitudes. Other studies should examine the relationship of assessment results to instructional practices or examine long-term changes in the interpretations of test scores within a state.

The extent of these consequences on validity are not apparent from Miller's (1999) study. Surveys such as these are not strong enough by themselves to show the degree of influence on score interpretation or use. Teacher surveys can only suggest how consequences affect score interpretation and use. For example, it is clear that teachers are aligning instruction with state-mandated PAs. However, instructional alignment with an assessment, or teaching the test, can take many forms (Mehrens & Kaminski, 1989; Miller & Seraphine, 1993). Different forms of instructional alignment result in different score interpretations. Similarly, professional development can vary substantially in form and content, and it has only an indirect effect on score interpretation and use. Future studies should examine the interpretations and uses of test results within the context in which the consequence occurred.

Conclusions and Recommendations

There are many unresolved validity issues with PAs. Considerable care and study have been devoted to specifying the content and scoring procedures of PAs and their generalizability across scorers; however, many areas have a small knowledge base, and more research is needed. Some pressing issues in PA validation are:

1. The large examinee-by-task variance component seems to limit the generalizability of PAs unless a large number of tasks is used. Given the time needed to complete a single complex performance-based task, more tasks in the assessment are often problematic. More study is needed to understand the boundaries of the interpretations and/or ways of reducing the interaction of examinees with tasks.
2. Consequences will continue to be important for PAs. They have been adopted in many states due to the negative consequences of MC assessments on instruction, combined with the presumed positive consequences of PAs on instruction and learning (e.g., curricular and instructional changes). These should be documented in conjunction with their effects on score interpretation and use (e.g., interviews and focus groups with key stakeholders). In addition, research should continue to examine and document any negative unintended consequences (e.g., bias).
3. Different forms of assessment are routinely being used. Limited evidence suggests that generalization across different forms of assessment might not be comparable, which means that two types of assessments cannot be easily combined. Moreover, interpretations might need to be constrained to a form of assessment, just as interpretations are constrained to a single content area. Future research should explore this hypothesis, including testing and/or documentation of this effect, and examining appropriate methods of combining assessments for more generalized interpretations.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington DC: Author.

Aschbaker, P. R. (1991). Performance assessment: State activity, interest, and concerns. *Applied Measurement in Education, 4,* 275–288.

Blank, R. K. (1989). *Development of a 50-state system of education indicators: Issues of design, implementation, and use.* Washington DC: Council of Chief State School Officers.

Blank, R. K., & Engler, P. (1992). *Has science and mathematics education improved since "A nation at risk"?* Washington DC: Council of Chief State School Officers.

Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement: Issues and Practice, 14 (4),* 21–24.

Brennan, R. L. (1992). *Elements of generalizability theory.* Iowa City IA: American College Testing.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement, 14,* 385–396.

Brennan, R. L. (1996). Generalizability of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 19–58). Washington DC: National Center for Education Statistics. (NCES 96-802)

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24,* 339–353.

Burton, E. (1998). *An investigation of the school-level generalizability of performance assessment results.* Unpublished doctoral dissertation, University of Colorado, Boulder.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Candell, G. L. & Ercikan, K. (1994). On the generalizability of school-level performance assessment scores. *International Journal of Educational Research, 21,* 267–278.

Chudowsky, N., & Behuniak, P. (1998). Using focus groups to examine the consequential aspect of validity. *Educational Measurement: Issues and Practice, 17 (4),* 28–38.

Crocker, L. (1997). Assessing content representativeness of performance assessment exercises. *Applied Measurement in Education, 10,* 83–95.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington DC: American Council on Education.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57,* 373–399.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 621–694). Washington DC: American Council on Education.

Haertel, E. H., & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 59–78). Washington DC: National Center for Education Statistics. (NCES 96-802)

Harrison, J. M. (1998). *A comparison of strategies for estimating internal consistency on tests with missing scores.* Unpublished master's thesis, University of Florida, Gainesville.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6,* 125–160.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16 (2),* 14–16.

Linn, R. L., & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (pp. 84–103). Chicago: University of Chicago Press.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20,* 15–21.

Linn, R. L., & Herman, J. L. (1997). *Standards-led assessment: Technical and policy issues in measuring school and student progress* (CSE Technical Report No. 426). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

McLaughlin, M. W., & Shepard, L. A. (1995). *Improving education through standards-based reform.* A report by the National Academy of Educa-

tion Panel on Standards-Based Education Reform. Stanford CA: National Academy of Education.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16 (2),* 16–18.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice, 8 (1),* 14–22.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Messick, S. (1996). Validity in performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment*. Washington DC: National Center for Education Statistics. (NCES 96-802)

Miller, M. D. (1998). *Generalizability of performance-based assessments.* Washington DC: Council of the Chief State School Officers.

Miller, M. D. (1999). *Teacher uses and perceptions of the impact of statewide performance-based assessments.* Washington DC: Council of the Chief State School Officers.

Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Measurement: Issues and Practice, 12 (3),* 9–15.

Miller, M. D., & Seraphine, A. E. (1993). Can test scores remain authentic when teaching to the test? *Educational Assessment, 1,* 119–129.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher, 18 (1),* 16–25.

Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice, 16 (2),* 9–13.

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

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park CA: Sage.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership, 46 (7),* 4–9.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405–450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16 (2),* 5–8, 13, 24.

Stecher, B. M., Klein, S. P., Solano-Flores, G. McCaffrey, D., Robyn, A., Shavelson, R. J., & Haertel, E. (2000). The effects of content, format and inquiry level on science performance assessment scores. *Applied Measurement in Education, 13,* 139–160.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 46 (7),* 41–47.

Wise, L. L., Hauser, R. M., Mitchell, K. J., & Feuer, M. J. (1998). *Evaluation of the voluntary national tests: Phase I.* Washington DC: National Academy Press.

Yen, W. M. (1997). The technical quality of performance assessments: Standard errors of percents of students reaching standard. *Educational Measurement: Issues and Practice, 16 (3),* 5–15.

Author's Address

Send requests for reprints or further information to M. David Miller, 1403 Norman Hall, P.O. Box 117047, University of Florida, Gainesville FL 32611, U.S.A. Email: dmiller@coe.ufl.edu.