

Validity Evidence for Assessments¹

Suzanne Lane
University of Pittsburgh

As described in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) validity refers to “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). The validation process, therefore, involves the accumulation of evidence to support the proposed test score interpretations and uses. The process of accumulating evidence to support the validity of test score interpretations starts prior to the development of an assessment (AERA, APA, & NCME, 1999). As Messick (1989) has stated, validation is a continuous process and begins with a construct in search of appropriate assessment instruments and procedures. Sources of validity evidence have been identified by a number of theorists and researchers (e.g., AERA, APA, & NCME, 1999; Cronbach, 1988; Kane, Crooks, & Cohen, 1999; Kane, 1992; Linn, Baker, & Dunbar, 1991; Linn, 1993; Messick, 1989).

Haertel (1999) has pointed out, however, that the process of accumulating validity evidence is more than a checklist procedure. The validation process involves the development and evaluation of a coherent validity argument for and against proposed test score interpretations and uses (Haertel, 1999; Messick, 1989; Cronbach, 1988; Kane, 1992; Kane, Crooks, & Cohen, 1999). Each inference in the validity argument is based on an assumption or proposition that requires support. Setting forth a validity argument allows for the accumulation of evidence not only for, but also against intended test score interpretations. As stated by Messick (1992), the validation process involves accumulating evidence for and examining potential threats to the validity of test score interpretation. Moreover, Kane, Crooks, and Cohen (1999) argue that the “... the most attention should be given to the weakest part of the interpretative argument because the overall argument is only as strong as its weakest link” (p. 15).

The use of educational assessments at the local, state, and national levels has become more prevalent within the last two decades. Most states have implemented assessment programs that are being used for high-stakes purposes such as holding schools accountable to improved instruction and student learning as well as for grade promotion and certification. Many state assessment programs depend in part on performance-based tasks (e.g., Kentucky, Maryland, and Massachusetts) which are considered critical tools in the educational reform movement (Linn, 1993). Kane, Crooks, and Cohen (1999) stated that although all assessments, including performance-based assessments, should be evaluated based on the same validation criteria, “... the emphasis given to each step in

¹ This paper is based on a presentation given at the 1999 Edward F. Reidy Interactive Lecture Series sponsored by The National Center for the Improvement of Educational Assessment, Inc., held October 14-15, 1999, Providence, RI.

the interpretation may be different for different kinds of assessments and interpretations” (p. 15). For performance assessments, they suggested that special consideration should be given to the generalizability of the results over tasks, raters, and occasions.

To determine what validity evidence is necessary, analysts should delineate a set of propositions that would support the proposed interpretations for the particular purpose of testing. Evidence should then be collected to support each proposition. As an example, for a state high school certification test developed to determine whether students mastered the state content standards, examples of relevant propositions include:

- (a) the test content is representative of the state content standards,
- (b) the test scores can generalize to other relevant sets of items,
- (c) the test scores are not unduly high or low due to irrelevant constructs being measured, and
- (d) the students’ curriculum and instruction afforded them the opportunity to attain the state content standards. (See AERA, APA, NCME (1999) for other examples.)

The purpose of the present paper is to describe the sources of evidence that can be accumulated to help support or refute a validity argument. The evidence, however, should not be collected in a piecemeal fashion but should be continuously evaluated as an integrated set to determine the extent to which the validity argument is supported. As stated by the *Standards for Educational and Psychological Testing* (1999), “... a sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses (p. 17).” This paper discusses evidence related to test content, response processes, internal structure, external structure, generalizability of test score interpretations, and consequences of test score interpretation and use. It should be noted, however, this paper does not provide exhaustive coverage of validity evidence for assessments. For example, evidence related to scale construction, such as evidence for the model used in evaluating the psychometric properties of items and evidence for the linking/equating procedures used, is not addressed in this paper (see for example, AERA, APA, NCME (1999) for an overview). Finally, Haertel (1999) challenges us to think about other meaningful, yet less traditional forms of validity evidence to help support test score interpretation and use.

As previously indicated, the importance of each type of evidence depends on the intended interpretations and uses of assessment results. Consider the example of the state high school certification assessment that was developed to determine the level at which students mastered the state content standards. For such an assessment, accumulating two forms of evidence would be pertinent: the extent to which the state assessment reflects the state’s content standards and the extent to which the curriculum offered to students reflected the content standards. It should be noted the weight given to discussing any one source of evidence in this paper is not necessarily related to the importance of that evidence.

Test Content Evidence

Messick (1989) has argued that construct theory serves as a guide to the development of an assessment and provides a rational basis for specifying features of items, rubrics, and scoring procedures as well as for expecting certain empirical evidence (e.g., degree of homogeneity of item responses and relationships between scores with other measures). Further, the validity of score interpretations is dependent on the fidelity of the construct that is measured by the test and the resulting test scores (Messick, 1989). As Messick (1989) has stated, an assessment is an imperfect measure of a construct, in part due to the extent to which it underrepresents the construct domain (i.e., the assessment is too narrow). The degree to which one can generalize from performance on an assessment to the larger construct domain depends on whether the breadth of the content represented in the assessment reflects the breadth of the defined construct domain. The development of an assessment, therefore, begins with a clear definition of the construct and the intended purpose of the assessment. The relationship between the test content and the construct the test is intended to measure provides an important source of validity evidence. As indicated in the *Standards for Educational and Psychological Testing* (1999), test content includes the format, wording, and context of items as well as scoring procedures, guidelines, and rubrics.

Representativeness is of particular importance when using performance-based assessments. Kane, Crooks, and Cohen (1999) highlight an important paradox when using performance-based assessments; that is, there is a trade-off between the congruency between the test content and the construct domain, and the extent to which generalizations from a small sample of tasks to the construct domain can be made. High fidelity items tend to be time-consuming resulting in a small number of items on the assessment. This in turn undermines the generalizations of the scores to the construct domain.

An assessment is also an imperfect measure of a construct to the extent that it measures one or more irrelevant constructs in addition to measuring the intended construct (Messick, 1989). Messick (1989) has identified two sources of construct-irrelevant variance: construct-irrelevant difficulty and construct- irrelevant easiness. Examples of potential sources of construct-irrelevant difficulty that leads to scores that are unduly low are the level of reading comprehension required by a mathematics assessment, and unfamiliarity with item wording, context, and format. Potential sources of construct-irrelevant easiness that lead to scores that are unduly high are flaws in item format and wording. As an example, Figure 1 provides an example of a QCAI mathematics task for which a large percentage of students in the pilot testing interpreted the task in a manner inconsistent with what was expected (Lane & Parke, 1992). The students' interpretations were just as valid given the way in which the task was worded. However, these students were responding to a simple and uninteresting mathematical problem.

Figure 1

The Robinson family owns a company which makes cleaning supplies. They need to buy a new machine which makes buckets.

They see the three advertisements below.

The Galaxy Bucket Machine



Makes 21 Buckets in just 15 minutes!

The Industrial Bucket Machine



Makes 82 Buckets every hour!

The Heavy Duty Bucket Machine



Makes 44 Buckets in only 30 minutes!

The family wants to buy the machine which makes buckets the fastest.

A. Which machine do you think they should buy?

Answer: _____

B. Why do you think they should buy this machine?

Item and response formats that are inconsistent with the construct domain may also be a source of construct-irrelevant variance (i.e., irrelevant method variance Messick (1989)). An example would be an assessment that uses only multiple-choice items for evaluating students' writing proficiency.

Examples of Validation Studies

This section describes studies that can provide test content evidence for the validity of educational assessments.

- Evaluate the extent to which the test specifications are aligned to the construct domain. This can be accomplished by determining whether the content and processes reflected in the test specification are relevant or not relevant to the construct definition. Also, such an analysis provides additional evidence

regarding the appropriateness of the construct definition and may lead to further delineation of the construct domain.

- Evaluate the extent to which the assessment items and rubrics are representative of the construct domain.
- Conduct fairness reviews of the assessment items and rubrics using a panel of content experts who are knowledgeable about fairness issues related to relevant groups of examinees. Task wording, format, and context needs to be considered in light of differing cultural and linguistic backgrounds and prior experiences of students. Differences due to familiarity with task wording, format, and context can affect the validity of interpretations of assessment results. As Duran (1989) has indicated, in order for students to demonstrate their maximum performance, they need to understand the various modes of thinking and reasoning that are expected, and the ways in which language is used in the assessment context. Fairness reviews are conducted to evaluate the extent to which item content (a) may portray stereotypes and/or be offensive to one or more groups and (b) is the source of irrelevant variance (i.e., construct-irrelevant difficulty and construct-irrelevant easiness). Irrelevant item difficulty may be due to lack of prior knowledge of material in reading passages, or undue complexity in reading material for a mathematics assessment. Irrelevant item easiness may be due to aspects of the item or test format that allows some individuals to respond correctly in ways irrelevant to the construct being assessed.
- Conduct content reviews of the assessment items and rubrics using a panel of content experts. Pilot studies of items and rubrics should also be conducted to determine whether the items and rubrics are functioning as intended and to evaluate the extent to which item content and rubric criteria are the sources of irrelevant variance.
- Evaluate procedures for administration and scoring such as the appropriateness of instructions to examinees, time limit for the assessment, and training of raters.
- For tests used in making decisions about promotion and graduation, evaluate the extent to which students have had the opportunity to learn the content of the assessment.

Examples of Research Studies Providing Test Content Evidence

Sireci (1998) discusses traditional approaches and a new procedure for obtaining validity evidence based on test content. He argues for the use of both item-objective congruence and relevance procedures as well as newer approaches such as the use of multidimensional scaling analysis of item-similarity ratings (see Sireci & Geisinger (1995) for a description of this procedure). The latter approach requires judges to rate the similarities among all pairs of test items with regard to the content and processes measured by the items, without the use of the test specification. The data are then subjected to a multidimensional scaling analysis resulting in a visual display of the judges' perceptions of the similarity among items. The groupings of the items in the visual display are used to evaluate how well the test measures the construct domain. He

further provides a set of guidelines for accumulating validity evidence when item-similarity ratings are used. Some of these guidelines are also relevant for the traditional approaches including: select competent and representative judges, select representative samples of items, use rating scales with more than 5 points and use even-numbered scales, familiarize the judges with the items and rating procedures, make the rating task as simple as possible, provide frequent breaks, provide incentives to the judges, evaluate judges' understandings of the rating procedures, and evaluate the criteria employed by the judges (Sireci, 1998).

When evaluating procedures for scoring, the time limit for the assessment is one aspect that should be examined. By statistically comparing hierarchical graded IRT models using two groups of students who received differing amounts of administration time, my colleagues and I identified two constructed-response items on a mathematics performance assessment that were speeded (Lane, Stone, Ankenmann, & Liu, 1995). The results suggested that students who had more time to work on the assessment tended to perform better because of the strategy that they employed. For example, for one of the items a trial-and-error approach was used by many of the students. Although this approach was considered appropriate, it may have required more time than other strategic approaches resulting in lower scores for students receiving less test administration time.

Scoring rubrics used to evaluate student performance can affect the validity of test score interpretations in that they may include irrelevant criteria or fail to include important, relevant criteria (Kane, Crooks, and Cohen, 1999). Further, materials that accompany items and quality control procedures, such as evaluating raters' consistency and accuracy in assigning scores, may be inadequate. Taylor (1998) examined the impact of using three different scoring methods (i.e., holistic-scoring, trait scoring, and item-by-item scoring or step-by-step scoring) on the validity of score interpretations to a mathematics assessment. The results indicated that each scoring method assessed somewhat different characteristics of students' performances. The results also suggested that constructs such as "mathematical communication" and "concepts and procedures" were interwoven based on a factor analysis of the data. As Taylor (1998) stated, the latter result "...suggests that test development strategies that attempt to classify items as measuring either mathematical concepts or mathematical communication (or reasoning or problem solving) may not be valid" (p. 217).

Response Process Evidence

An examination of the extent to which the cognitive skills and processes identified in the test developer's defined construct domain are elicited from the examinees when taking the assessment provides validity evidence. The specification and analysis of the cognitive requirements of the tasks as well as the analyses of processes, strategies, and knowledge underlying task performance provide such evidence. For example, student responses to constructed-response tasks have the potential to provide concrete traces of their processes and strategies. As indicated by Glaser (1990), judgements regarding the

cognitive significance of an assessment begin with an analysis of the cognitive requirements of the tasks as well as the ways in which students attempt to solve them.

Messick (1989) discusses several techniques that can be used to analyze the processes and strategies underlying task performance. One method is protocol analysis, in which students think aloud as they solve problems, or describe retrospectively their solution processes. Another method is the analyses of students' rationales for their answers and ways of responding (analyses of reasons). A third method is the analysis of errors, in which the researcher draws inferences about processes from incorrect procedures, concepts, or representations of the problems. These types of logical analyses provide validity evidence to support or refute the use of items in an assessment.

As the *Standards for Educational and Psychological Testing* (1999) state, the extent to which the process of raters is consistent with intended score interpretations is a source of validity evidence. As indicated by Messick (1992), features of the rubrics should be reflected in the scores assigned by raters to help ensure the validity of the test score interpretations. With more states using constructed-response items and performance-based tasks in their assessment programs, evidence is needed to ensure that raters are interpreting and using the scoring criteria accurately when assigning scores to students' performances.

Examples of Validation Studies

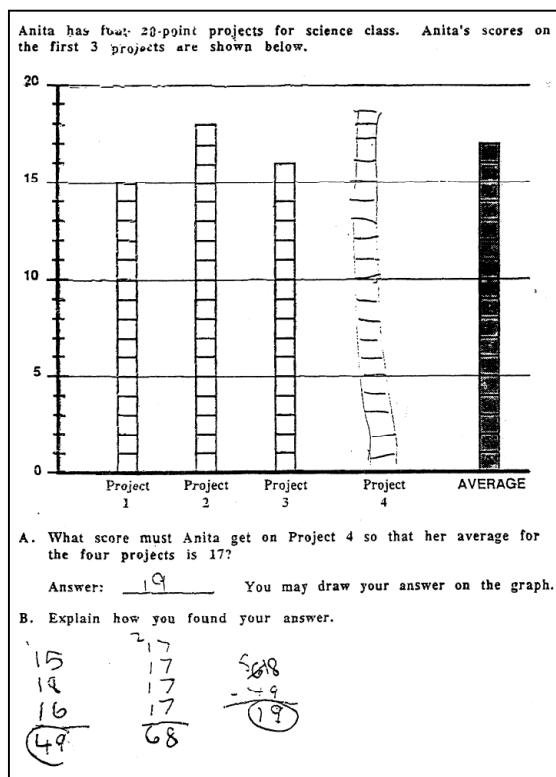
This section describes some validation studies that can provide response process evidence for educational assessments.

- Describe and document how each item assesses the cognitive skills and processes it was intended to assess.
- Evaluate the extent to which assessment items are eliciting the cognitive skills and processes that were intended, and are not eliciting the skills and processes that were not intended using protocol analyses (think aloud procedures or 'cognitive labs') or analyses of student written explanations for constructed-response items.
- Evaluate the processes employed by various subgroups to determine the extent to which irrelevant variables are affecting their performance differentially. As stated by the *Standards for Educational and Psychological Testing* (1999), response process evidence "...can contribute to questions about differences in meaning or interpretation of test scores across relevant subgroups of examinees (p.12).
- For constructed-response items and performance-based assessments, evaluate the extent to which raters apply the scoring criteria appropriately and are not influenced by irrelevant factors in scoring students' performances (e.g., interpretation of the criteria is accurate, personal biases do not affect scoring).

Examples of Research Studies Providing Response Process Evidence

To evaluate the extent to which QCAI mathematics constructed-response items elicited the processes and reasoning they were intended to elicit, students' justifications, solution strategies and errors were analyzed during pilot testing as well as during the operational administration. As an example, the Average Task in Figure 2 was designed to elicit a variety of solution strategies that may be used by students from different instructional programs (Lane & Silver, 1994). As indicated in the figure, the results suggested that students were using a variety of strategies. The strategies included a visual strategy (or "leveling-off" strategy), solving an equation, trial and error, and a "balancing" strategy. The analyses also indicated that students became more proficient over time in applying strategic processes when solving this task. Thus, evidence was also provided to support the assumption that the task was sensitive in measuring improvements in students' use of strategic processes (see Magone, Cai, Silver, and Wang (1994) for additional response process evidence of the QCAI items).

Figure 2



As another example, Hamilton, Nussbaum, & Snow (1997) used interviews with students in conjunction with statistical analyses to help define the constructs underlying the NELS:88 tests. The interview data helped clarify the dimensions identified by factor

analytic procedures. A study conducted by Paulsen, Best, Levine, Milne, and Ferrara (1999), using cognitive labs, revealed that the most common problem on extended constructed-response items identified by students was unclear language or contexts. These types of problems could interfere with the extent to which the cognitive skills and processes identified in the test developer's defined domain are elicited from the students.

Research has indicated that raters may not always use the features delineated in rubrics when scoring student papers (e.g., Pomplun, Capps, and Sundbye, 1998; Rafoth & Rubin, 1984). Pomplun, Capps, and Sundbye (1998) examined the extent to which raters were using and interpreting rubrics accurately when assigning scores to student responses on constructed-response mathematics and reading items on Kansas's state assessment. They examined rubric-related and rubric-unrelated features used by teachers when assigning scores to determine the extent to which the validity of the test score interpretations was undermined by construct-irrelevant variance. Their results indicated that for mathematics the correct answer and quality of reasoning which were rubric-related features were influential in the holistic scores assigned by raters. The length of the response, a rubric-unrelated feature, was also influential in the scores assigned for one grade of the mathematics assessment and for all grades of the reading assessment. Pomplun and his colleagues suggested that this may be due to teachers using personal constructs (Huot, 1990) when rating student responses. Citing Brookhart (1993), they further suggested that this influence of response length might be because when assigning classroom grades teachers tend to consider effort, progress and completion.

Internal Structure Evidence

When examining the internal structure of an assessment, the extent to which the individual items and the assessment, itself, measure the intended construct(s) is of primary interest. As indicated by the *Standards for Educational and Psychological Testing* (1999), evidence based on the internal structure of the assessment indicates "... 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).

Internal structure evidence can also be obtained by examining whether items behave differently for subgroups of students of approximately equal ability (i.e., differential item functioning (DIF) studies can be conducted). DIF refers to items that do not function the same after groups have been matched with respect to the attribute being measured (Holland & Thayer, 1986). Differential item functioning, however, is a statistical finding and may not necessarily warrant removal of items that are flagged as DIF when the content quality of the assessment may be jeopardized (Angoff, 1993; Doolittle & Clearly, 1987). Rather items that exhibit DIF may have implications for curriculum and instructional changes (Harris & Carlton, 1989). For performance-based tasks, in addition to conducting statistical analyses to examine DIF, logical analyses of student responses and/or performances can be conducted to evaluate plausible reasons for DIF that are more directly related to differences in students' ways of thinking and responding.

Examples of Validation Studies

The following are validation studies that can provide internal structure evidence for educational assessments.

- For assessments intended to measure one construct, evaluate the extent to which each assessment item differentiates students along this single construct and the extent to which items are interrelated (using, for example, exploratory factor analyses or confirmatory factor analyses).
- For assessments intended to measure more than one construct, the extent to which resulting scores are meaningfully different can be evaluated.
- Conduct differential item functioning (DIF) studies to examine the extent to which examinees with the same ability, but from different groups, are responding similarly to a given item. DIF studies can indicate unintended or intended multidimensionality of test data.
- Analyze student responses to evaluate potential reasons for DIF-detected constructed response items (e.g., differences in strategies adopted by groups).

Examples of Research Studies Providing Internal Structure Evidence

Kupermintz and his colleagues (Kupermintz, Ennis, Hamilton, Talbert, & Snow (1995); Kupermintz & Snow, 1997) examined the internal structure of the National Educational Longitudinal Study of 1988 (NELS: 88) mathematics tests for 8th, 10th, and 12th grade students. In general, a two factor model (mathematical reasoning and mathematical knowledge) fit the 8th, 10th, and two forms of the 12th grade test questioning the validity of the interpretations based on a total score. Furthermore, the analyses revealed that one form of the 12th grade assessment had a “...much more complex pattern highlighting several more specialized aspects of performance within the mathematics domain” (Kupermintz & Snow, 1977, p. 132). The factor analyses of the NELS:88 science test also indicated multidimensionality and the need to use multiple scores rather than to make interpretations based on a total score (Nussbaum, Hamilton, & Snow, 1997).

DIF analyses can be used to examine the extent to which items may have differential validity for subgroups of students. A study conducted by my colleagues and me examined gender-related DIF for QCAI mathematics constructed-response items (Lane, Wang, & Magone, 1996). To complement the statistical analysis, logical analyses of middle-school student responses to the flagged mathematics items were conducted to determine potential reasons for DIF. The logical analyses of a few of the DIF items that favored female students over matched male students indicated that females were more explicit in showing their solution strategies and were more likely to provide conceptual explanations, in that, females tended to map their numerical answer back to the problem context. In fact, one of these items assessed the concept of ratio and proportions, which typically favors males. However, similar to other research, male students as compared to matched female students performed better on a geometry item and a more complex ratio

item. The results of the study suggested that some features that have been associated with DIF for multiple-choice items may not hold when the assessment includes constructed-response items and performance-based tasks.

Garner and Engelhard (1999) examined DIF on a mathematics high school graduation test and concluded that DIF is linked not only to the item content but also item format. They found that all of the constructed-response items that were flagged for DIF favored female college students. In the Lane et al. study, four of the six constructed-response items that were flagged for DIF favored middle-school female students.

External Structure

The relationships between scores on an assessment and other measures provide additional validity evidence. As the *Standards for Educational and Psychological Testing* (1999) indicate, “... Evidence based on relationships with other variables addresses questions about the degree to which these relationships are consistent with the construct underlying the proposed interpretations (p. 13)”. Validity evidence of this nature has been categorized as convergent, discriminant, and criterion-related evidence. Convergent evidence is provided by relationships between test scores and other measures intended to assess similar constructs. Discriminant evidence is provided by relationships among test scores and other measures intended to assess different constructs. Criterion-related evidence, either predictive or concurrent, is provided by relationships between test scores and examinees’ performance on a criterion measure (Cronbach, 1971, Messick, 1989).

Examples of Validation Studies

The following are validation studies that can provide external structure evidence for educational assessments.

- Evaluate the relationship among student scores on the assessment with variables that are intended to measure a similar construct or to be related. As an example, the relationship between scores from a high school certification exam and other measures can be evaluated. Other measures may include other district- or state-adopted tests, school grades, teacher ratings, NAEP, SAT, ACT, and Advanced Placement Tests.
- Evaluate the relationship among student scores on the assessment with variables that are intended to measure a different construct (e.g., relationship between a mathematics achievement test and reading comprehension).
- For high school certification exams, evaluate the extent to which the test scores predict relevant criterion performance as compared to other variables such as high-school QPA. Measures of criterion performance may include enrollment in college, college grades, patterns of college courses, job performance, and vocational school performance.

- Examine the relevancy and quality of the criterion, including the reliability and validity of the criterion scores. Practical problems of availability and convenience, however, need to be considered when selecting a criterion measure.

Examples of Research Studies Providing External Structure Evidence

Studies have been conducted to examine the relationship between the assessment that was used for the Kentucky Instructional Results Information System² (KIRIS; Kentucky Department of Education, 1997) and other measures including the Armed Service Vocational Aptitude Battery (ASVAB) and the ACT (Hoffman, 1998, Wise, 1997). As an example, Wise (1997) obtained convergent and divergent correlations between the scores from the ASVAB and KIRIS that suggested KIRIS measured its intended constructs. For example, the correlations between the ASVAB and KIRIS math measures were between .67 and .73 and the correlations between the ASVAB and KIRIS reading measures were between .54 and .56. It should be noted that the assessments used different item formats (multiple-choice vs open-response). For another example of external validity evidence for a state assessment see Public Schools of North Carolina (1996).

Generalizability Evidence

Messick (1989) has indicated the need for the “systematic appraisal of context effects in score interpretations, especially the degree of generalizability across different population groups, different ecological settings, different time periods, and different task domains or subdomains” (p.56). Empirical evidence can be obtained to determine the extent to which the score interpretations for an assessment can generalize to other population groups, to other situations or settings, to other time periods, and to other tasks representative of the construct domain.

Research examining whether the internal structure of an assessment is similar across various population groups, such as Mexican-American, African-American and Caucasians, can provide some validity evidence for population generalizability³. A study examining the degree to which the processes and strategies elicited by students who receive testing accommodations are congruent with the intended processes and strategies can provide validity evidence for ecological generalizability. Cross-sectional data as well as longitudinal data can provide evidence for temporal generalizability (see for example, Koretz & Baron (1998)). Further, the extent to which the test score interpretations can be generalized to the construct domain can provide evidence for task generalizability. For performance-based tasks and constructed-response items, error due to raters can also affect the generalizability of score interpretations.

² It should be noted that KIRIS has been replaced by another assessment system. However, similar analyses are planned for the new assessment.

³ It should be noted the terms population generalizability, ecological generalizability, temporal generalizability, and task generalizability were used by Messick (1989); however, the way in which they are used in the present paper is not entirely consistent with Messick (1989).

Examples of Validation Studies

The following are validation studies that can provide evidence on the generalizability of the score interpretations.

- Evaluate the extent to which the assessment has the same meaning across groups. For example, evaluate the extent to which the internal and external structure of the assessment is similar across relevant groups of examinees.
- Evaluate the appropriateness of accommodations provided to students; in particular, examine the extent to which the measured construct is similar to that of the general student population. As the *Standards for Educational and Psychological Testing* (1999) indicate, “... the purpose of accommodations or modifications is to minimize the impact of test-taker attributes that are not relevant to the construct that is the primary focus of the assessment” (p. 101).
- Evaluate changes in student performance on the assessment over time at each grade level.
- Evaluate the extent to which score interpretations can generalize across items, raters, and occasions.

Examples of Research Studies Providing Generalizability Evidence

With the increased use of performance-based assessments and constructed-response items, the majority of the empirical work has examined the extent to which test score interpretations can be generalized to the broader construct domain. Evidence pertaining to the generalizability of the scores to the broader construct domain has focused in part on the intertask consistency which is an essential piece of validity evidence for the use of an assessment and score interpretation (Dunbar, Koretz, & Hoover, 1991; Messick, 1989). The intertask relationships in writing, math, and science have indicated that the generalizability of individual-level scores derived from assessments consisting of relatively small number of performance-based tasks is questionable (e.g., Hieronymus & Hoover, 1987; Shavelson, Baxter, & Pine, 1991; Lane, Liu, Ankenmann, & Stone, 1996; McBee & Barnes, 1998). This is not unreasonable given the nature of the constructs being assessed and the variety of task formats, but this lack of generalizability affects the validity of score interpretations.

Cronbach, Linn, Brennan, and Haertel (1997) have argued, however, that task-sampling variability is confounded with occasion-sampling variability because students typically sit for the assessment on only one occasion. Shavelson, Ruiz-Primo, and Wiley (1999) provided support for this argument and concluded that both the person-by-task interaction and the person-by-task-by-occasion interaction were responsible for the large task-sampling variability. This highlights the need to include occasion as a facet in generalizability studies. As stated by Kane, Crooks, and Cohen (1999), “Large values for the variance components associated with any source of error can undermine inferences from observed scores to universe scores and, therefore, undermine the interpretative argument as a whole” (p. 10).

DeMars (2000) investigated the generalizability of score interpretations when testing was conducted under different contexts. More specifically, she examined how scores changed on science and math sections of Michigan's high school proficiency test when the potential consequences of the test changed. The low-stakes test administration was the final pilot administration, and the high-stakes condition was the operational test administration that was used for endorsing state diplomas. Her results indicated that students performed better on the high-stakes administration as compared to the low stakes administration; however, the difference was greater for constructed-response items than multiple-choice items. These findings imply that validity of score interpretations can be jeopardized to varying degrees depending on the stakes of the test as well as the item formats included in the test. As DeMars (2000) indicated, these findings also imply that equating test forms on pilot data may lead to inaccurate score interpretations, particularly for tests composed of constructed-response items. Consistent with previous findings reported in this paper, her results indicated that males tended to outperform females on multiple-choice items, however, females tended to outperform males on constructed-response items. These results suggest that the assessment may measure something irrelevant to the construct of interest, or that the construct may need to be reconceptualized.

Zuriff (2000) investigated whether extra examination time for college students with learning disabilities unduly favored their test performance in relation to students with no identified learning disabilities. His analysis of five studies revealed that there is only weak support to the proposition that college students without learning disabilities would not benefit from the extra examination time. This proposition is based on the assumption that students without learning disabilities already perform at their maximum potential under the given time conditions. Zuriff concluded that additional empirical evidence is needed to support the practice of providing extra test administration time for college students with learning disabilities. It should be reiterated, however, that the sample used in these studies consisted of college students. See the *Standards for Educational and Psychological Testing* (1999) for a comprehensive discussion on the validity of test score interpretations for students with learning disabilities and students with limited English proficiency.

Consequential Evidence

In 1997 and 1998, two issues of *Educational Measurement: Issues and Practice* (1997, 1998) focused primarily on the consequential aspects of validity. One of the issues highlights a debate among researchers on whether the consequences of test use and score interpretation are an integral part of validity research as espoused by Messick (1989, 1992). Shepard (1997) and Linn (1997, 1998) argued that the consequences of test use and score interpretation are an integral aspect of validity. In contrast, Popham (1997) asserted that such consequences of assessment programs do not fall within the realm of validity. These researchers, however, agree that consideration of and attention to consequences of educational assessments is essential. Finally, the *Standards for*

Educational and Psychological Testing (1999) state that evidence about consequences is relevant to validity when it can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components.

Researchers have also pointed out that rigorous evidence for the consequences of an assessment has yet to be obtained (Kane, Khattri, Reeve, & Adamson, 1997; Mehrens, 1998). As indicated by Mehrens (1998), causative inferences cannot be drawn from the evidence that has been collected. Reckase (1998) highlighted the problems in obtaining causal evidence for the expected consequences using the ACT as an example. However, he indicated that, although there is no causal evidence for the consequences of the ACT, there is empirical evidence for the relationship between the test scores and grades in entry-level courses. Thus, in line with Mehrens' reasoning, it may be reasonable to evaluate the extent to which empirical evidence suggests positive and/or negative consequences.

Frederiksen and Collins (1989) proposed that assessments have “systemic validity” if they encourage behaviors on the part of teachers and students that promote the learning of valuable skills and knowledge, and allow for issues of transparency and openness, that is access to the criteria for evaluating performance. The accumulation of evidence of interpretations of assessment results by teachers, students, administrators, and policymakers, as well as the actions they take as a consequence, should be undertaken for educational assessment programs.

Examples of Validation Studies

The following are validation studies that can provide consequential evidence for assessments.

- For an assessment that is intended to improve instruction, examine the extent to which intended and unintended changes in classroom instructional practices occur, such as instructional time spent on the content and processes assessed by the assessment versus other content and process areas. This may be accomplished by using focus groups, interviews, questionnaires, classroom artifacts, and classroom observations.
- For tests used to make graduation decisions, examine the impact of the assessment outcomes on student career and college decision-making as well as academic and career opportunities afforded to students.
- Evaluate the extent to which various groups of users (e.g., students, teachers, principals, general public, media) interpret assessment results appropriately. Questions that can be posed with regard to the interpretation of assessment results are: What appropriate and inappropriate interpretations do users have of assessment results when various types of scores are reported (e.g., percentiles, proficiency levels)? What appropriate and inappropriate interpretations do users have of assessment results reported in different formats (e.g., tabular, graphic, interpretative text)?

Examples of Research Studies Providing Consequential Evidence

Pomplun (1997) demonstrates a method for investigating consequential evidence of validity for the Kansas state assessment that was developed to facilitate change in instructional practices. Using path model analyses with the data source being teacher questionnaires, his results indicated that teacher- reported professional activities and attitudes toward the state assessment, especially toward the scoring rubric, were related directly to changes in instructional practices. Stone and Lane (2000) examined the relationship between changes in school performance on the Maryland School Performance Assessment Program (MSPAP; MSDE (1995)) and teacher, student, and school variables using growth models. More specifically, they examined the relationship between changes in MSPAP scores for schools and classroom instruction and assessment practices, student motivation, students' and teachers' beliefs about and attitude towards MSPAP, and school characteristics. The results indicated that teacher reported instruction-related variables explained differences in performance on MSPAP across five subject areas, and for some subject areas, explained differences in rates of change in MSPAP performance over time. In addition, teacher perceived impact of MSPAP on instruction and assessment practices was also found to explain differences in MSPAP performance levels or rates of change over time across the subject areas.

Wainer, Hambleton, and Meara (1999) examined preference for and appropriate interpretation of NAEP data displays by educational policy makers in state departments of education. Their results indicated that although these users' may not prefer a particular display, they may be more accurate in extracting information from that display rather than a "preferred" display. This result suggests that studies need to examine not only user preference, but also the accuracy in interpreting data displays prior to their operational release.

Concluding Remark

This paper discusses some sources of validity evidence for test score interpretations and uses; however, it does not provide a comprehensive treatment of the topic (see the *Standards for Educational and Psychological Testing* (1999) for a more comprehensive discussion).

References

- AERA, APA, & NCME (1999). *Standards for Psychological Testing*. Washington, DC: American Psychological Association.
- Brookhart, S. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement*, 30, 123-142.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., p. 443-507). Washington, DC: American Council on Education.
- Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test Validity* (pp. 3-17), Hillsdale, NJ: Erlbaum.
- 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(3), 373-399.
- DeMars, C. E. (2000) Test stakes and item format interactions. *Applied Measurement in Education*, 13(1), 55-78.
- Duran, R. P. (1989). Testing of linguistic minorities. In R.L. Linn, (Ed.), *Educational Measurement* (3rd ed.) (p. 573-388). New York: American Council on Education.
- Garner, M. & Engelhard, Jr. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, 12(1), 29-52.
- Haertel, E.H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, 18(4), 5-9.
- Hamilton, L.S., Nussbaum, M., & Snow, R. E. Interview procedures for validating science assessments. *Applied Measurement in Education*, 10(2), 181-200.
- Hoffman, R. G. (1998). *Relationships among KIRIS Open-Response Assessment, ACT Scores, and Students' Self Reported High School Grades*. (HumRRO Report FR-WATSD-98-27). Radcliff, KY: Human Resources Research Organization.
- Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, 60, 237-263.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5-17.

- Kane, M. B., Khattri, N., Reeve, A. L., & Adamson, R.J. (1997). *Assessment of Student Performance*. Washington, D.C.: Studies of Educational Reform, Office of Educational Research and Improvement, U.S. Department of Education.
- Koretz, D. M., & Barron, S. I. (1998). *The Validity of Gains in Scores on Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND.
- Kuppermintz, H., Ennis, M. M., Hamilton, L. S., Talbert, J. E., & Snow, R. E. (1995). Enhancing the validity and usefulness of large-scale educational assessments: I. NELS:88 Mathematics Achievement. *American Educational Research Journal*, 32, p.525-554.
- Kuppermintz, H. & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: III. NELS:88 Mathematics Achievement to 12th Grade. *American Educational Research Journal*, 34(1), p. 124-150.
- Lane, S. & Parke, C. S. (April 1992). Principles for developing performance assessments: An example of their implementation. Paper presented at the Annual Meeting of the American Educational Research Association.
- Lane, S. & Silver, E. A. (April 1994). Examining students' capacities for mathematical thinking and reasoning in the QUASAR project. Paper presented at the Annual Meeting of the American Educational Research Association.
- Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1995). Examination of the assumptions and properties of the graded item response model: An example using a mathematics performance assessment. *Applied Measurement in Education*, 8(4), 313-340.
- Lane, S., Wang, N., Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practice*, 15(4), 21-27, 31.
- Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, 15(1), 1-16.
- 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. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice*, 17(2), 28-30.
- Linn, R.L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15-21.
- Magone, M. E., Cai, J. Silver, E. A., & Wang, N. (1994). Validating the cognitive complexity and content quality of a mathematics performance assessment. *International Journal of Educational Research*, 21(3), 317-340.

Maryland State Board of Education (1995). *Maryland School Performance Report: State and School Systems*. Baltimore, MD.

McBee, M.M & Barnes, L. L. B. (1998). The generalizability of a performance assessment measuring achievement in eighth-grade mathematics. *Applied Measurement in Education*, 11(2), 179-194.

Mehrens, W.A. (1998). Consequences of Assessment: What is the Evidence? *Evaluation Policy Analysis Archives*, 6(13).

Messick, S. (1992). *The interplay of evidence and consequences in the validation of performance assessments* (ETS RR-92-39). Princeton, NJ: Educational Testing Service.

Messick, S. (1989). Validity. In R.L. (Ed.), *Educational Measurement* (3rd ed.) (p.3-104). New York: American Council on Education.

North Carolina State Board of Education (1996). *End-of-Grade Tests: Reading Comprehension and Mathematics*. Raleigh, North Carolina: Author.

Nussbaum, E. M., Hamilton, L. S., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: IV. NELS:88 Science Achievement to 12th Grade. *American Educational Research Journal*, 34(1), p. 151-173.

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

Poplum, M. (1997). State assessment and instructional change: A path model analyses. *Applied Measurement in Education*, 10(3), 217-234.

Poplum, M., Capps, L., & Sundbye, N. (1998). Criteria teachers use to score performance items. *Educational Assessment*, 5(2), 95-110.

Rafoth, B. A., & Rubin, D. L. (1984). The impact of content and mechanics on judgments of writing quality. *Written Communication*, 1,446-458.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice*, 17(2), 13-16.

Shavelson, R.J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement*, 36(1), p. 61-71.

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

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, 5(4), 299-321.

- Sireci, S. G. & Geisinger, K. F. (1995). Using subject matter experts to assess content representation: A MDS analysis. *Applied Psychological Measurement, 16*, 241-255.
- Stone, C. A. & Lane, S. (2000). MSPAP performance gains from 1993-98 and their relationship to “MSPAP impact” and school characteristic variables. Paper presented at the Annual Meeting of the National Council of Measurement in Education.
- Taylor, C. S. (1998). An investigation of scoring methods for mathematics performance-based assessments. *Educational Assessment, 5(3)*, 195-224.
- Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement, 36(4)*, 301-335.
- Wise, L. L. (1997). *Merging ASVAB and KIRIS On-Demand Scores: Report of Preliminary Results (LRS97-4)*. Frankfort, KY: Bureau of Learning Results Services, Kentucky Department of Education.
- Zuriff, G. E. (2000). Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. *Applied Measurement in Education, 13(1)*, 99-117.