Threats to the Valid Use of Assessments

TERRY J. CROOKS[1], MICHAEL T. KANE[2] & ALLAN S. COHEN[3]
[1]*Educational Assessment Research Unit, University of Otago, Box 56, Dunedin, New Zealand,* [2]*Department of Kinesiology, University of Wisconsin-Madison, 2000 Observatory Drive, Madison, WI 53706, USA &* [3]*Testing and Evaluation Services, University of Wisconsin-Madison, 1025 West Johnson Street, Madison, WI 53706, USA*

ABSTRACT *Validity is the most important quality of an assessment, but its evaluation is often neglected. The step-by-step approach suggested here provides structured guidance to validators of educational assessments. Assessment is depicted as a chain of eight linked stages: administration, scoring, aggregation, generalization, extrapolation, evaluation, decision and impact. Evaluating validity requires careful consideration of threats to validity associated with each link. Several threats are described and exemplified for each link. These sets of threats are intended to be illustrative rather than comprehensive. The chain model suggests that validity is limited by the weakest link, and that efforts to make other links particularly strong may be wasteful or even harmful. The chain model and list of threats is also shown to be valuable when planning assessments.*

Introduction

Validity is the most important consideration in the use of assessment procedures. The primacy of validity is enshrined in professional standards (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1985) and reaffirmed in most books and articles on assessment. Recent efforts to build a more coherent and unified view of validity have expanded its scope and further strengthened its importance (Cronbach, 1980, 1988; Messick, 1989, 1994; Linn *et al.*, 1991; Kane, 1992; Moss, 1992; Shepard, 1993; Linn, 1994). The breadth and centrality of validity, as now conceived, is clearly evident in Messick's recent definition:

> Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. (1989, p. 13)

As Linn (1994, p. 6) has noted, however, there is a difference between affirming the primacy of validity and acting upon it. In practice, validity has often received less attention than reliability or generalizability (Gipps, 1994). A major reason for this appears to have been the discrepancy between the algorithmic nature of procedures for

0969-594X/96/030265-21

estimating generalizability and the more open-ended nature of procedures for estimating validity. Generalizability estimation is based on mathematical procedures and numerical indices and is relatively easy to standardise, report and defend. Validity estimation, on the other hand, relies heavily on human judgement and is therefore harder to carry out, report and defend. Furthermore, it is always vulnerable to one new piece of negative evidence, which may undermine confidence in the assessment. It is probably no accident that the most commonly reported estimates of validity involve numerical correlations between different assessments (criterion-related validity coefficients), nor that construct validity evidence has been much more prominent in advocacy than in use.

Although the recent broadening and unifying of the concept of validity has been well-argued and soundly based, the very breadth and complexity of the concept makes it difficult to work with in practice and, therefore, threatens continuing neglect of validity in the monitoring of the quality of assessments. Validation will only flourish if new approaches are developed which help us to organise our thinking about important validation questions and to identify issues which need particularly close scrutiny (Shepard, 1993).

One such approach has been to identify sets of validity criteria which should be considered (see, for instance, Haertel, 1985; Cole & Moss, 1989; Frederiksen & Collins, 1989; Linn *et al.*, 1991; Messick, 1995). As Moss (1992) has noted in a review, some of these schemes relate to assessment in general, while some are concerned primarily with performance assessment. The sets of criteria proposed have proven helpful in identifying issues which deserve attention in validation, and in clarifying how specific concerns relate to the more global issues of construct under-representation and construct irrelevant variance (Messick, 1989, 1994).

Another approach has been presented by Kane (1992) and Shepard (1993), based on the concept of validity argument (Cronbach, 1988). They noted that interpretations of performances on assessments necessarily involve a linked series of inferences and assumptions. If these inferences and assumptions can be identified, their plausibility can be examined by logical and empirical means, and their importance can be debated. Both authors give examples of how their proposed approach to validation can be applied in practice. The approach is powerful and meaningful, but while it points out general directions for exploration, the validator is left with the major task of finding a suitable pathway for the validity argument.

In this paper, we suggest an approach that combines the virtues of a clearly defined set of validation criteria and the structure of an argument-based approach. Assessment is depicted as divided into eight conceptually distinct stages, with validation then based on careful scrutiny of each of these stages. The eight stages are likened to eight links of a chain, with weakness of any one link weakening the chain as a whole. Further guidance is offered to the validator through identification of several possible validity threats associated with each link. These lists of threats are intended to be illustrative rather than comprehensive.

By identifying the eight links to be considered, and listing a substantial number of validity threats associated with these links, our model complements the validity argument approach of Kane and Shepard. It suggests that it may be helpful to consider

the validity argument in eight sections (corresponding to the eight links), and it provides numerous examples of specific flaws which can occur in the interpretation and use of assessment data (the threats). Most of the threats discussed here have been identified by other researchers, but they have not been placed in such a structured model. The approaches suggested by Cole & Moss (1989) and Haertel (1985) are most similar, because they also have used time sequences of assessment processes as their organising schema.

Our approach can be applied to any assessment procedure. For simplicity in presenting and discussing threats, we have chosen threats and examples relating to assessments of student achievement. All eight links in the assessment chain deserve consideration in each case, although their relative influence on validity can be expected to vary in different cases (for instance, some links may be particularly important where assessment is mainly intended to be formative, but less so where it is mainly intended to be summative). Most of the listed threats will also deserve consideration in each case, but users of the approach should also attempt to identify any further threats which are associated with their particular assessment context.

We focus here primarily on the validation of existing assessment tasks and procedures. Accordingly, the first link in our model is the administration of the assessment tasks. Clearly, however, validation can only take place if the intended purposes of the assessment are well understood. The appropriateness of the assessment tasks and procedures to those purposes will be a central issue in evaluating the strength of each link in the assessment chain. While progress towards the intended purposes must be evaluated, so too must unintended side Some of the possible side effects are included among the listed threats.

Because our prime focus is on validation, the initial discussion of the model does not include any direct reference to the planning of the assessment and the development of the assessment tasks. The issues which should be addressed by developers are, however, extensively covered in the links included in the model, and explicit consideration of the links and the associated threats should be very helpful during the development process. Much of the evidence for validation should therefore be developed during the design of the assessment procedures and the development of the assessment tasks, but needs to be verified and supplemented from the users' perspective.

The model can also usefully be applied to the planning of assessments and the development of assessment tasks. For this purpose, the eight links need to be considered in reverse order, working backwards from the intended interpretation and use to ensure that all of the links in the chain of inference are sound.

Overview of the Validation Model

Our model of validation is based on Fig. 1, which depicts assessment as involving eight linked stages:

(1) *Administration* of assessment tasks to the student.
(2) *Scoring* of the student's performances on the tasks.

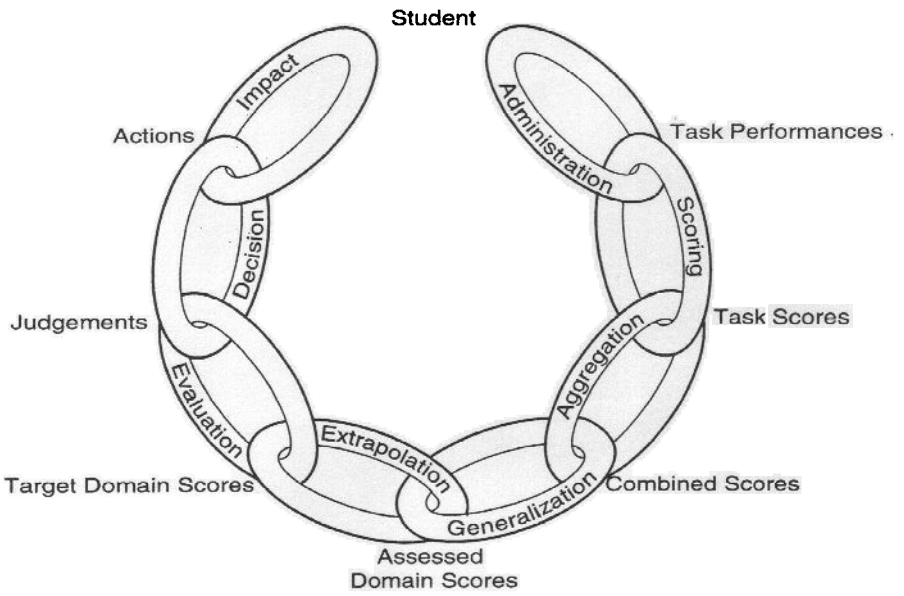

FIG. 1. A model of educational assessment for use in the validation and planning of assessments.

(3) *Aggregation* of the scores on individual tasks to produce one or more combined scores (total score or subscale scores).
(4) *Generalization* from the particular tasks included in a combined score to the whole domain of similar tasks (the *assessed domain*).
(5) *Extrapolation* from the assessed domain to a *target domain* containing all tasks relevant to the proposed interpretation.
(6) *Evaluation* of the student's performance, forming judgements.
(7) *Decision* on actions to be taken in light of the judgements.
(8) *Impact* on the student and other participants arising from the assessment processes, interpretations, and decisions.

The administration link is included because task performances can be greatly influenced by the procedures followed in presenting and administering the tasks.

Scoring, aggregation and generalization are often merged into a single link (reliability or generalizability) in other models of assessment and validation, but there are substantial advantages to be realised in evaluating them separately. The threats to each of these links are conceptually distinct, even if there may be interactions across the links. For instance, aggregation deals with the interpretability of composite scores derived from particular assessment tasks, while generalization involves inferences from the composite scores to the domains from which the tasks have been selected.

A similar case can be made for distinguishing between the extrapolation, evaluation and decision links. Merging them, as is commonly done, obscures the inferential leaps between performance scores in a domain of tasks, interpretations of the performance (commonly using psychological constructs), and decisions on actions to be taken

(which require judgements about the most appropriate and effective ways to use the assessment information).

The impact link reminds us that thorough consideration of the consequences of assessment processes is an essential component of a comprehensive review of validity evidence (Cronbach, 1988; Messick, 1989).

The importance of the eight links in the model can be illustrated by mentioning for each link just one example of the threats to validity associated with that link. Validity may be seriously undermined if one or more of the following circumstances apply: some students receive inappropriate help with the tasks (administration link); scoring of some or all of the tasks emphasises unimportant but easily rated aspects of the performances (scoring link); scores for tasks which are very heterogeneous are added together (aggregation link); few tasks are used, so a small sample of performance is obtained (generalization link); no tasks are included from some substantial sections of the target domain (extrapolation link); performance is interpreted using construct language without supporting evidence (evaluation link); the standards used in making decisions are inappropriately high or low (decision link); or actions resulting from the assessment undermine the educational progress of many of the students (impact link).

Validation requires careful consideration of the strength of each of the eight links. We suggest that anyone using the model for validation begin by considering the administration link and then move clockwise around the other seven links (Fig. 1), observing the assessment process in action and evaluating the threats to validity associated with each of the links.

This approach applies to all assessments of student achievement, but the relative importance of each link (i.e. its potential to compromise validity) will vary depending on how the assessment is used. For instance, for a classroom-based assessment intended solely for diagnostic and formative purposes, the aggregation, generalization and extrapolation links may be somewhat less important than other links. Alternatively, for a summative assessment addressing broad constructs, those same three links will be very important.

In the next section of this paper, we detail some threats to validity associated with each of the eight links (Table I). Readers are reminded that the lists of threats are not claimed to be comprehensive. This section is followed by a discussion of three important implications which can be derived from the eight-link model and by some suggestions for the practical use of the model in validating assessments. Finally, we briefly demonstrate that the model provides useful, systematic guidance for developers of assessment procedures, applying it to some major issues in the design of national systems for monitoring educational outcomes.

Threats Associated with Each Link

Administration

The administration of the assessment tasks to students is the first step in the assessment process. Close examination of task administration is therefore the first link in the validation chain. The circumstances under which student performances are obtained

TABLE I. Some threats associated with each of the eight links

Link	Threat
Administration	Low motivation Assessment anxiety Inappropriate assessment conditions Task or response not communicated
Scoring	Scoring fails to capture important qualities of task performance Undue emphasis on some criteria, forms or styles of response Lack of intra-rater or inter-rater consistency Scoring too analytic Scoring too holistic
Aggregation	Aggregated tasks too diverse Inappropriate weights given to different aspects of performance
Generalization	Conditions of assessment too variable Inconsistency in scoring criteria for different tasks Too few tasks
Extrapolation	Conditions of assessment too constrained Parts of the target domain not assessed or given little weight
Evaluation	Poor grasp of assessment information and its limitations Inadequately supported construct interpretation Biased interpretation or explanation
Decision	Inappropriate standards Poor pedagogical decisions
Impact	Positive consequences not achieved Serious negative impact occurs

can have major implications for the validity of the interpretations and actions resulting from an assessment. Any of the four threats discussed below can seriously reduce the validity of score interpretations and consequent decisions for some or all students.

Low motivation. If students are not motivated to do well on assessment tasks and therefore do not put maximum effort into completion of the tasks, it will be misleading to interpret their performance as indicative of their knowledge, skill or ability. Low motivation can arise from a variety of circumstances. For instance, students may perceive the results of the assessment to be of little importance to them. Low motivation can also occur when students believe they have very little chance of success on the tasks or when they perceive the assessment tasks to be artificial and irrelevant to their lives. Authenticity (Wiggins, 1993) or meaningfulness (Linn *et al.*, 1991) of tasks can help to counter low motivation.

Assessment anxiety. The antithesis of low motivation is assessment anxiety, which occurs for some students when motivation is high. Such anxiety undermines student performance, giving a misleading picture of what the student might be able to do under

less anxious conditions (Hill & Wigfield, 1984). High-stakes assessment leading to important decisions about individual students is particularly likely to provoke assessment anxiety. Only some students are seriously affected; others may be aroused by the same conditions to peak levels of preparation, concentration and performance. Hill & Wigfield (1984) provide several suggestions for minimising the threat to validity posed by assessment anxiety.

Inappropriate assessment conditions. Student performance may be inappropriately low because proper procedures for administering the tasks were not followed. For instance, the administrator may have conducted the assessment under poor environmental conditions, allowed too little time, or failed to read the instructions fully and allow time for students to complete practice examples.

Alternatively student performance may be inappropriately high. Students may have received coaching on the specific assessment tasks, had access to resources not normally permitted, received direct help during the assessment, or presented someone else's work as their own. The administrator may have allowed more than the standard time to complete the assessment. The dangers of positive bias are greatest with high-stakes assessment, where teachers, students or both are under pressure to score well.

Task or response not communicated. Student performance may be misinterpreted as inability to carry out the task when in fact the task has not been properly understood, or when the central elements of the task have been performed but students have been unable to communicate their result. For example, the capabilities of students with physical disabilities may be misrepresented because their disabilities interfere with their ability to perceive the task requirements or to perform the physical actions needed to demonstrate their knowledge, ability or skills. A little less obviously, a student may be described as poor at mathematics when the real problem is that the student was unable to read the instructions for the mathematics tasks, but would have been able to perform the tasks if they had been presented in a different way (perhaps orally, or in a different language).

The problem of poor task communication is, however, much more widespread than the special nature of these two examples might suggest. Confusing instructions or poorly designed response arrangements can undermine the performance of all students taking a test, leading to inappropriate interpretations or decisions.

Scoring

Once the student performances have been gathered, the next stage of the assessment process is the scoring of each student's performance on each task. Close examination of the scoring process is therefore the second link in the validation chain. We discuss here five threats associated with scoring which can reduce the validity of score interpretations and consequent decisions, for some or all students.

Scoring fails to capture important qualities of task performance. Unrecognised errors in answer keys or scoring rubrics can deprive students of appropriate credit for their efforts. More commonly, a scoring rubric (explicit or covert) may take account of only some qualities of the performance, inappropriately ignoring other important qualities. For instance, the scoring of oral reading could legitimately take into account such aspects as accuracy of decoding, self-correction, fluency and expressiveness, all of which may be important in an overall rating of oral reading performance. If only one or two of the aspects are used in the scoring and the interpretation of the scores is not correspondingly restricted, the validity of the interpretation is undermined. Attempts to increase rater agreement by using more objective scoring criteria will often lead to a narrowing of the factors included in the scoring, thereby increasing the risk posed by this threat to validity.

Undue emphasis on some criteria, forms or styles of response. The validity of assessment interpretations or decisions may be restricted if scorers favour some styles of response over others. For instance, scorers of history tasks who place considerable weight on the correctness of students' written expression (spelling, grammar, etc.) might be doing an injustice to students whose knowledge of history and skill in historical analysis are strong, but who are poorly equipped to write well in English.

Lack of intra-rater or inter-rater consistency. If scorers are not consistent with themselves or each other in the performance aspects they consider, the standards they set, or the marks they award, the validity of assessment interpretations or decisions is threatened. It is desirable to reduce the extent of such inconsistency, but not at the expense of eliminating or reducing the weight given to important aspects of task performance which can only be assessed though professional judgement (and which may therefore tend to be judged variably by different scorers).

Scoring too analytic. If the scoring process provides for separate scores on many different aspects of performance, there is a danger that the richness of the whole performance will not be captured. For instance, students writing a critical analysis of a research article may receive separate scores or ratings for the total number of strengths and weaknesses they correctly identified in the article, for the correctness of their grammar, syntax and spelling, and (negatively) for the number of areas incorrectly identified as strengths or weaknesses. Although these appear to be appropriate factors to identify, these scores may not adequately reward a student who concisely, forcefully and convincingly identified the most significant strengths and weaknesses. A more global scoring procedure would allow such clarity of insight and purpose to be properly rewarded.

Scoring too holistic. The use of holistic approaches to scoring is often wasteful of the information available, reducing the validity of the assessment. For instance, scoring a substantial student project by awarding a single grade and providing no additional information on the project's strengths and weaknesses lowers the validity of the assessment for formative purposes. It would be much better to provide, in addition to

the grade, ratings of several key aspects of the performance and comments on how weak areas could have been improved.

Messick (1994, pp. 19–21) has provided an excellent analysis of the virtues and perils of assessing what he calls complex and decomposed skills, demonstrating clearly that the extent of aggregation both in scoring of individual tasks and in combining scores from different tasks should be determined largely by the purpose of the assessment.

Aggregation

When all tasks have been scored, scores from individual tasks or components of tasks can be aggregated to produce subscale or total scores. Close examination of aggregation across tasks is the third link in the validation chain. We discuss here two threats to validity which are associated with aggregation of task scores.

Aggregated tasks too diverse. If too wide a range of tasks is included in an aggregated score, many of the correlations among tasks will be low, reducing the coherence of the aggregated score. This lack of homogeneity will limit the generalizability, interpretability and usefulness of the aggregated score. Under these circumstances, it will often be desirable to group the tasks into more coherent subsets and compute aggregate scores for each subset. If needed, a total score can also be computed. What is appropriate will depend largely on the proposed uses of the assessment information.

For instance, science is a broad curriculum area and not very hierarchical, raising doubts about the wisdom of aggregating across diverse aspects of biology, chemistry, physics, astronomy and geology. Consider the three tasks investigated by Shavelson *et al.* (1993), which involved student explorations of the absorbency of paper towels, the environmental conditions preferred by sow bugs, and electrical components hidden in boxes. Given the diversity of topics and the extent to which the electricity task favoured students with prior knowledge, it may not be surprising that relatively low inter-task correlations were found. Perhaps the correlations among tasks would have been substantially higher if all the tasks involved investigations relating to electricity, or all involved the behaviour of insects.

Inappropriate weights given to different aspects of performance. For the aggregated score to be most meaningful, the weights given to different tasks should reflect the relative importance of the tasks within the assessed domain. Inappropriate weighting can result from poor balance in the number of tasks in different areas, from the scoring procedures for different tasks, or from differences in the score variance for different tasks. For instance, an assessment being used to rank a group of students may involve multiple-choice items and an essay, with the intention that the abilities assessed by the two types of tasks be equally weighted. Because of the relative difficulties of the tasks and the particular procedures adopted in scoring the tasks, however, the ranking could be determined predominantly by performance on either the multiple-choice items or the essay.

Generalization

When aggregated scores for students on a set of tasks have been obtained, the next step in interpreting the results is to generalize from the specific occasion, tasks and procedures used in the assessment to the associated *assessed domain* (defined below). A close examination of generalizability (reliability) is therefore the fourth link in the validation chain.

A particular set of tasks and the conditions under which they are administered, can be viewed as a random sample from a much larger collection of tasks and conditions which could equally well have been used in the assessment. We are calling this larger collection the assessed domain. If an assessment process involves two or more clusters of tasks (or aspects of tasks) and a score for each cluster, then there is a different assessed domain for each of those clusters.

Generalizability refers to the accuracy of generalizing from a student's aggregated score to his or her universe score in the corresponding assessed domain (the *assessed domain* score). The assessed domain score is the expected (average) score across all permissible assessments that could be drawn from the assessed domain. Such assessments would sample performance on the full range of tasks in the domain and under the varied conditions (such as occasion, administrator, location and time allowed) which are possible.

We discuss here three threats to validity associated with generalizability. They are all quite closely related to some threats mentioned in earlier links, but have different emphases and implications in this link.

Conditions of assessment too variable. Student performance on the assessment may vary substantially depending on such variables as the time allowed for completion of the tasks, the time of day when the assessment is administered, the task formats used to assess students' abilities, and the approaches adopted by different administrators. If this is the case, failure to control these factors restricts the generalizability of the assessments. By standardising such factors, it may be possible to increase generalizability substantially. It must be recognised, however, that such standardising may substantially narrow the assessed domain, so that subsequent interpretations and decisions based on the students' performances must either be more constrained or involve greater extrapolation (Kane, 1982).

Inconsistency in scoring criteria for different tasks. An important factor in generalizability is the level of correlation among scores on the different tasks included in the assessment. There are a number of possible reasons for low inter-task correlations, one of which is inconsistency in the scoring criteria for the different tasks. If these scoring criteria can be made more similar, the generalizability should improve. The risk in doing so, however, is that the assessed domain will be narrowed as a result of the more standardised scoring criteria. The implications of such standardisation have been widely discussed in relation to performance assessments, for which some scholars have favoured task-specific scoring rubrics and others have favoured generic rubrics. As Messick (1994) has pointed out, more standardised criteria and greater emphasis on

generalizability are appropriate where a broad construct-centred interpretation is the main goal, but not when considerable attention is being given to performance on individual tasks.

Too few tasks. Dependability is limited by the size of the sample of student behaviour used in the assessment. As the size of the sample drawn from the assessed domain decreases, the dependability of generalizations to that domain also tends to decrease.

Extrapolation

An assessed domain is usually only a subset of the corresponding domain that will be used in interpreting the assessments (the *target domain*). While a set of tasks and conditions of assessment can be viewed as a random sample from their assessed domain, they are often a systematically biased sample from the target domain because some categories of tasks or conditions of assessment in the target domain have been excluded from the assessed domain.

The desired interpretation involves an extrapolation from the universe score for the assessed domain to the universe score for the target domain (the *target domain score*). A close examination of the extent and trustworthiness of this extrapolation is the fifth link in the validation chain. We discuss here two threats to the validity of assessment interpretations and decisions associated with the extrapolation process.

Conditions of assessment too constrained. If the assessment has been conducted under constrained conditions, narrower than the range of conditions permitted in the target domain, it may be misleading to treat the universe score in the assessed domain as equivalent to the universe score in the target domain. For instance, if all tasks have been presented as multiple-choice items, the results may not give a sound indication of performance on similar curriculum goals but assessed through different task formats. Similarly, fixed time limits for task completion can significantly influence performance but may not be present or required in the target domain.

Parts of the target domain not assessed or given little weight. Substantial differences between the educational goals which are included in the assessed domain and those in the target domain raise doubts about the wisdom of extrapolating from the universe score for the assessed domain to the universe score for the target domain. The degree of risk to the validity of the extrapolation process varies inversely with the degree to which the assessed domain covers the target domain. The risk will be exacerbated if performance on the included tasks is weakly correlated with performance on the excluded tasks.

The importance of this threat to validity, which Messick (1989, 1994) has called construct under-representation, cannot be overstated. It is interesting to note that three of the eight validity criteria for performance assessments described by Linn *et al.* (1991) are related to this one threat. The labels Linn *et al.* used for these three criteria (content coverage, content quality and cognitive complexity) highlight important issues which all need to be addressed.

Failure to assess some significant parts of the target domain may occur because the developer of the assessment has not recognised the existence or importance of those parts of the domain, and therefore has not devoted sufficient effort to developing assessment tasks in these areas (content coverage). Alternatively, the developer may have constructed tasks intended to assess performance in the areas, but not succeeded in focusing the tasks well enough to achieve the intended purpose (content, quality, cognitive complexity).

Evaluation

The sixth step of the assessment process is evaluation, leading from target domain scores to judgements about the merit (and perhaps the strengths and weaknesses) of the student's performance. The goal of the evaluation step is to answer the question, 'What do the target domain scores mean?' If the assessment includes several dimensions, a profile of scores from several different target domains will need to be evaluated. We discuss here three threats to validity which can arise in the evaluation process.

Poor grasp of assessment information and its limitations. No matter how sound the assessment has been up to this point, its validity can be seriously undermined if the person evaluating the assessment information does not properly understand the information or the limitations arising from its selective nature and the particular arrangements used to collect it. This can lead to inappropriate judgements.

For example, a student may decide that she is not good at history on the basis of assessments of her history work by a particular teacher whose assessment (and teaching) placed great stress on recalling factual details, such as dates of events. She might have reached a very different conclusion with a teacher whose assessments emphasised students' skills in examining evidence, discussing alternative interpretations, and writing coherently about these. Failure to recognise the limited nature of the assessments can lead to a potentially serious misinterpretation.

The danger of misinterpretation is probably greatest where the person interpreting the assessment information has not been involved in designing the assessment. No matter how thorough and accurate the interpretive guidelines available (in a test manual or elsewhere), they do not help if they are not carefully read and understood. This issue deserves particular consideration when classroom teachers are making use of the results of standardised tests, or reporting and interpreting the scores to students or parents.

Inadequately supported construct interpretation. Almost all evaluations of assessment performances involve some use of social, psychological or educational constructs, and some of these uses of constructs require very large inferential leaps from performance to construct. For instance, the proposition that creativity can be judged by administering tasks which require students to describe unusual uses for common objects involves a very substantial inferential leap, and one which has been widely disputed. As the magnitude of such an inferential leap increases, the potential risk to

the validity of the interpretation also increases and more substantial evidence is required to support its validity.

Interpretations of assessment information often involve more subtle use of constructs. As an example, consider a student who performs well on a series of science laboratory reports and tests. A comment that the student performed very well does not involve a substantial construct interpretation, but a comment that the student is 'good at science' clearly does, because it invokes the construct of ability in science. Such uses of construct language slip in easily as words are chosen to describe performance, and therefore require careful scrutiny.

Biased interpretation or explanation. A person interpreting an assessment score rarely does so in a vacuum. In addition to the information provided by the current assessment, the person is likely to be aware of the student's performance on earlier assessments and of other information about the student. This additional knowledge can be very helpful, but it can also be dangerous. At best, it results in a more accurate and useful interpretation, and therefore in better decisions. At worst, it results in seriously biased interpretations of the assessment information.

As an example of the beneficial effects of additional information, consider a student who performed poorly on one particular assessment, despite a consistently good record on other tasks in that subject. The teacher notes that the student had been ill during the days leading up to the assessment, and decides that little weight should be given to the performance on this particular occasion.

A related but contrasting example involves the halo effect and an inappropriate positive bias. As in the example above, a student performs poorly on one assessment but has a past record of good performances. The teacher makes allowances for the student, perhaps even scoring the work more positively on a second consideration. If the student really does not understand the material covered by the assessment, or is losing motivation to study the subject, the teacher's action may not only distort the assessment of the student but also delay the provision of needed help.

A final example involves negative bias. A student obtains an unexpectedly high but legitimate score on a standardised assessment, The teacher or counsellor, however, treats this score as an anomaly to be ignored rather than as important information which requires revised judgements about the student's capabilities, or at the very least some further investigation.

Decision

With judgements made, the next step of the assessment process is to decide what actions to take as a result of the judgements. These actions may be as limited as deciding to give no feedback and to allow a student to proceed with coursework, or as major as deciding that a student should not be admitted to a college to which he or she has applied. Examining the merit of decisions taken is the seventh link of the validation chain. A good decision will be consistent with the information on which it is based (which often includes information obtained on earlier occasions, before the current

assessment), and will also result in generally beneficial consequences for students and other participants in assessment processes. We have identified two threats to validity associated with this link.

Inappropriate standards. Standards play a key role in many assessment decisions. Standards may be public and explicit, in which case they define what grade will be awarded for a given assessment score, what score or pattern of scores will be sufficient for admission to a particular college, or what score will result in a decision requiring a student to relearn a unit of work and demonstrate improved understanding by taking a further assessment. Other assessment situations involve more informal standards, perhaps existing only in the mind of an assessor. Inappropriate standards can undermine the validity of decisions based on assessment information, and therefore of the assessment as a whole.

As an example, consider the classic predictive validity situation in which a cut-score on an assessment is being used to determine who gets into an educational programme and who does not. If there are many more applicants than places available, the cut-score serves a simple rationing purpose and the only validity evidence required is that scores on the assessment are substantially correlated with performance in the educational programme and the cut-score awards admission to the desired number of candidates. If, however, the number of places in the programme is not restricted, the validity requirements are more demanding. A substantial correlation is still required, but it is now necessary to look closely at the effect of the choice of cut-score on both the proportion of admitted students who will fail and the proportion of non-admitted students who would have succeeded. The validator should be able to demonstrate in light of this analysis that a good choice of cut-score has been made.

Poor pedagogical decisions. It may seem strange that we are identifying pedagogical decisions as threats to the validity of an assessment. However, deciding what action to take arising from an assessment (such as what feedback to give to a student) is a crucial part of the overall assessment process. This involves very important pedagogical issues, and not merely interpretation of the assessment information. Pedagogical decisions play a major role in determining the impact of the assessment, and therefore directly influence the assessment's validity.

Consider, for instance, a teacher who decides that a student has performed very well overall on a set of tasks, but who chooses to provide feedback to the student focused heavily on the defects in the student's work. By doing so, the teacher may undermine the validity of the assessment, at the very least giving the student a misleading picture of the teacher's overall judgement, and perhaps also damaging the student's motivation in the subject.

To complicate the picture further, it will usually be helpful to take into account individual differences between students. Students who are confident that they are capable of performing well are likely to be less at risk from feedback which concentrates

on their errors than students who regularly make many errors and lack confidence in their ability to do better.

Because decisions made by the assessor usually have direct impacts on students and other participants in the assessment, there is a close relationship between the decision link and the last link in the assessment chain, the impact link. There is inevitably some overlap and interaction between these links, as there is between earlier links in the chain.

Impact

The final step of the assessment model is qualitatively different from the previous seven, which deal with identifiable stages of the assessment process. The eighth step looks at the impact of the assessment on students and other participants in the assessment process. Much of what Messick (1989, p. 20) has called 'the consequential basis of validity' needs to be considered here. In addition to the direct and indirect impact of assessment decisions on participants, it is important to consider the effects of experiencing the whole process of assessment. There is extensive evidence (e.g. Crooks, 1988; Madaus, 1988) that both large scale and classroom assessment can have major effects on participants.

No matter how technically sound the first seven steps of the assessment may be, the impact of the assessment may call the validity of the entire assessment into question. Accordingly, the eighth link in the validation chain involves close scrutiny of the consequences of assessment processes and actions. We discuss here two threats to validity associated with this link.

Positive consequences not achieved. The effort involved in the assessment process can only be justified if the assessment leads to worthwhile benefits for students or other stakeholders. For students, benefits could include academic credit, appropriate placement in educational programmes, helpful feedback to improve learning, enhanced motivation, or greater confidence in skills and future performance. Benefits for other participants might include trustworthy identification of students who will perform well in an educational programme or workplace, or feedback to teachers which will help them refocus and improve their teaching.

Serious negative impact occurs. The actions arising from assessment decisions often have important negative consequences. Examples of potential negative consequences for students include reduced motivation, reduced self-efficacy, increased anxiety, exclusion from further learning opportunities, and focusing on factual learning at the expense of higher cognitive level outcomes (Crooks, 1988). Validity is also reduced if the assessment processes are perceived to be unfair, involve more than temporary and manageable stress for participants, or substantially damage relationships among participants.

Where such negative effects can be anticipated or recognised, even for a minority of participants, strong justification should be required if the assessment process is to

continue without significant modification. The justification should include counter-balancing evidence of positive effects, as well as evidence that negative effects have been minimised.

Implications of the Model

The Importance of the Weakest Link

The strength of a chain depends on its weakest link. By using a chain model for evaluating threats to the validity of an assessment, we are emphasising that validity is constrained by the strength of the weakest of the eight links in the validity chain. Some threats can be examined using quantitative techniques and thus are somewhat more straightforward to study, leading to a temptation to focus on strengthening these links while ignoring other links that are more difficult to study. There is, however, nothing to be gained from ensuring that some links are particularly strong unless the weakest link can also be strengthened.

One link which has often been omitted from discussions of validity is the first one (*administration*). Here the raw information on student performance is collected. The results are used in the following steps, where the observations are analysed, interpreted and used for decision making. Just as computer users are reminded that 'garbage in' leads to 'garbage out', assessment users need to be aware of the importance of task administration to the validity of the resulting interpretations and actions.

Through the inclusion of the *impact* link, the model acknowledges that the consequences of an assessment are an integral part of the validity of the process. An essential part of the validation of an assessment process is an examination of the extent to which the assessment achieves the purposes for which it was intended, and the extent to which both intended and unintended effects of the assessment are positive or negative for participants. Issues relating to the ethics and justice of the consequences of the assessment must not be ignored.

The Role of Reliability/generalizability in Validity

Linn (1994, p. 13) has echoed other scholars in stating that 'generalizability and other technical characteristics such as comparability derive their importance from the contribution they make to an overall evaluation of validity'. Our model is consistent with this view. Generalizability (reliability) issues can be identified in two of the eight links of the validation chain. Issues of intra-rater and inter-rater consistency are included in the scoring link, while issues of generalization from task scores to the assessed domain universe score are the central focus of the generalization link.

The model is consistent with the traditional view that some degree of generalizability is essential for validity, and that generalizability establishes an upper limit for validity. If either or both of the scoring and generalization links are weak, the chain as a whole is weak, and the assessment process has limited validity.

The chain model clearly illustrates that high generalizability is not sufficient for high validity. Generalizability is associated with only two of the eight links. No matter how strong those links are, the chain will be weak if any of the remaining links is weak.

Kane (1982) has demonstrated, using generalizability theory, that some strategies for enhancing generalizability can undermine validity. In particular, greater standardisation of assessment processes can be counterproductive. An examination of relationships between some of the threats to validity we have listed lends support to Kane's conclusion. Some possible strategies for increasing generalizability suggested by our listed threats are: removing sources of inconsistency in the scoring of tasks, greater standardisation of the conditions of assessment, using a narrower range of task formats, using a narrower range of task content, using more consistent scoring criteria across tasks, and more extensive aggregation of tasks. However, these strategies increase other threats to validity associated with both the scoring and extrapolation links, and probably also increase some threats in the administration, interpretation and impact links. Without doubt, undue emphasis on obtaining high generalizability can undermine validity.

Moss (1994) has argued that in some circumstances it may be possible to have validity without generalizability. Our model does not support that view. The model is, however, consistent with her view that the relative importance of different threats to validity, and in particular generalizability threats, vary greatly depending on the purposes of different assessments.

Different Assessment Purposes Imply Different Validity Emphases

The goal of achieving a strong assessment chain applies to all applications of assessment, and a strong chain requires that all links be considered for each application of assessment. Nevertheless, the level of risk to validity associated with each link varies markedly with different purposes for assessment. Therefore, the degree of concern about the different links should vary correspondingly.

Consider, for example, assessments which are intended to provide feedback to students on their performance on individual tasks. This purpose does not place emphasis on aggregation of tasks, generalization to a large domain of similar tasks, extrapolation to a broader domain, or a broad construct interpretation. It also does not usually require multiple raters or evidence of inter-rater agreement. However, the feedback will be of little value if the teacher is not reasonably consistent from day to day in the criteria sought and standards applied, or if the approach used in providing feedback does not help and motivate the students to improve their performance. Many of the threats to validity we have described are of limited or no relevance to this assessment purpose, and it is clear that the aggregation, generalization and extrapolation links are less susceptible to threats than are the other five links.

As a contrasting example, for a multiple-choice examination used to report the students' overall level of achievement at the end of a science course, the only threat in the scoring link is an error in the answer key, and the evaluation and decision links may also be of limited concern. However, issues related to the representativeness of sampling of the target domain and the generalizability of the scores are crucial, so that threats to the aggregation, generalization and extrapolation links must be very carefully

evaluated, and the administration and impact links also deserve careful consideration. Given the great disparity between the test format and the target domain, the extrapolation from the assessed domain score to the target domain score is likely to be suspect a priori and the threats to this link deserve special attention.

Practical Applications of the Model

Applying the Model to Validation

The primary purpose of the model is to guide and assist the validation of assessment procedures, interpretations and consequences. Its main advantage over other models of validation is that it provides more detailed guidance for the user through the eight-link structure and the identification of numerous threats to validity.

For any particular assessment, the first step in validation is to evaluate the importance of the eight links and the accompanying threats in light of the assessment's purpose. As previously discussed, different purposes imply substantially different emphases in validation, because the relative risks associated with each of the links and with the specific threats vary greatly with different assessment purposes.

The possible threats to validity associated with each of the eight links should be considered briefly, starting with the administration link and proceeding clockwise around Fig. 1. Threats not listed here should be considered alongside the threats we have listed. It will usually be reasonable to dismiss some threats as irrelevant or insignificant for the particular assessment and its purpose, while others will be judged as being potentially important.

When the potentially important threats have been identified, they can be evaluated individually. Both empirical and conceptual strategies will be needed. For instance, students can provide information (through interview or questionnaire) relevant to many of the threats in the administration and impact links. Statistical analyses will be appropriate for many of the threats in the scoring, aggregation and generalisation links. The remaining threats require conceptual analysis and professional judgement.

Because the strategies required are varied and can be quite complex, there is insufficient space to address them here in detail. The middle section of Messick's major review article (Messick, 1989) includes discussion of strategies for evaluating threats to validity, and Shepard (1993) presents four useful cases studies.

Validation will be easier if it has been planned for prior to the gathering of assessment information and if extra information has been gathered in conjunction with the assessment activities. For instance, evaluation of threats associated with the scoring link will be enhanced if samples of work have been routinely marked by more than one marker and if alternative marking schemes (e.g. both analytic and global) have been used for more complex tasks.

The final step in validation should be identification of the weakest links in the chain—the most important limiting factors in the overall validity of the assessment. This can only be done using professional judgement. The weakest links will be appropriate places to focus particular effort when trying to improve the assessment.

Applying the Model to Assessment Planning

Although our main purpose in this paper has been to show that the chain model can be a very useful tool for validating an existing assessment process, the eight-link model is also a valuable guide when designing assessment procedures. However, when planning an assessment the links should be considered in reverse order (i.e. moving counter-clockwise around Fig. 1), starting with the impact and decision links and ending with the administration link.

The purpose of the assessment must be clarified first, with careful consideration given to the actions expected to be taken based on the assessment as well as the desired impact both of those actions and of the overall assessment process. After this crucial first step, threats associated with each remaining link in the assessment chain can be considered systematically and the associated planning decisions made. Each decision will have important consequences for the validity of the assessment, and for the remaining development work.

In effect, then, the sequence of links and associated threats provides a structured guide which developers can follow to maximise the validity of their assessment. For example, the decision link directs attention to the form of reporting or feedback to be used, so that an appropriate impact can be achieved. For the evaluation link, it will be important to clarify whether the intended use involves a construct-centred interpretation or a domain- or task-centred interpretation. A key issue for the extrapolation link will be the extent to which practical factors (such as time constraints, cost, and student-safety concerns) prevent the full target domain from being sampled in the assessment, thus resulting in a narrower assessed domain. The generalisation link requires consideration of the homogeneity of the assessed domain and the number of tasks required for adequate generalisability. When the aggregation link is reached, the desirable extent of aggregation must be considered in the light of the intended purpose of the assessment and the breadth of the assessed domain. The scoring link includes consideration of procedures for achieving consistency in scoring while retaining emphasis on the most important qualities in the student performances. Finally, two important matters to consider within the administration link are how to ensure that working conditions are fair and that students are well motivated.

To illustrate further the usefulness of the model in the development of assessment procedures, we will briefly consider a few of the issues involved in the development of a system for monitoring educational outcomes on a national level. National monitoring has two main purposes:

- to monitor trends in performance across time, so that progress can be celebrated and decline or lack of progress recognized and addressed;
- to provide information that promotes productive debate about areas of relative strength and weakness across the curriculum.

The effectiveness of national monitoring for both of these purposes will be enhanced if the tasks used are seen by students and community to be assessing important and interesting educational outcomes. Such tasks are likely to address the more important aspects of the target domain, thus strengthening the extrapolation link, and are likely

to generate results that are worthy of scrutiny and debate, thus strengthening the evaluation, decision and impact links. They are likely to motivate students, thus strengthening the administrative link. Furthermore, the tasks can serve as models for assessment tasks to be developed by classroom teachers (impact link).

National monitoring cannot be fully effective if it focuses on only some areas of the curriculum (extrapolation link). Adopting a narrow focus may direct the efforts of teachers and the concerns of the community toward the areas assessed and away from a broader array of worthy goals (impact link). Assessing only some areas deprives the educational community of information about outcomes in other curriculum areas, thus limiting the debate about educational outcomes (evaluation, decision and impact links).

The goals of national monitoring can be satisfactorily achieved by generalising from a small sample of students to a description of the population. This approach has considerable advantages. For a given expenditure, much more comprehensive and rich data can be obtained from a small sample than from a full national population, thus supporting the extrapolation link without weakening the generalisation link. The sampling approach also lowers the stakes associated with the examination programme by limiting the possibility of reporting data on student or school performance, thus reducing incentives for teachers to distort the results by 'teaching to the tests' (impact link). There are many other issues to be addressed in the design of systems for national monitoring. Consideration of these issues can, in a similar way, be facilitated through systematic consideration of the eight links and their associated threats to validity.

Conclusion

The model presented here provides a framework for evaluating the validity of assessment uses and interpretations, and for efforts to build validity into assessments which are under development. The model identifies eight steps in assessment use and interpretation, and indicates some ways in which each step can go wrong. As noted earlier, the threats to validity listed under each link are intended as examples. They are not intended to form a checklist. The eight links represent issues that need to be addressed in any validation effort. Examining each link and looking for weaknesses in the chain of inference, including those arising from common specific threats, provides a systematic approach to validation.

References

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION, AND NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (1985) *Standards for Educational and Psychological Testing* (Washington, DC, American Psychological Association).

COLE, N.S. & MOSS, P.A. (1989) Bias in test use, in: R. L. LINN (Ed.) *Educational Measurement*, 3rd edn. pp. 443–507 (New York, American Council on Education/Macmillan)

CRONBACH, L.J. (1980) Validity on parole: how can we go straight?, in: *New Directions for Testing and Measurement: measuring achievement, Progress over a Decade, Proceedings of the 1979 ETS Invitational Conference*, pp. 99–108 (San Francisco, CA, Jossey Bass)

CRONBACH, L.J. (1988) Five perspectives on validity argument, in: H. WAINER & H. I. BRAUN (Eds). *Test Validity*, pp. 3–17 (Hillsdale, NJ, Erlbaum)

CROOKS, T.J. (1988) The impact of classroom evaluation practices on students, *Review of Educational Research*, 58, pp. 438–481.

FREDERICKSEN, J.R. & COLLINS, A. (1989) A systems approach to educational testing, *Educational Researcher*, 18, pp. 27–32.

GIPPS, C.V. (1994) *Beyond Testing: toward a theory of educational assessment* (London, Falmer Press).

HAERTEL, E. (1985) Construct validity and criterion-referenced testing, *Review of Educational Research*, 55, pp. 23–46.

HILL, K.T. & WIGFIELD, A. (1984) Test anxiety: a major educational problem and what can be done about it, *Elementary School Journal*, 85, pp. 105–126.

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

KANE, M.T. (1992) An argument-based approach to validity, *Psychological Bulletin*, 112, pp. 527–535.

LINN, R.L. (1994) Performance assessment: policy promises and technical measurement standards, *Educational Researcher*, 23, pp. 4–14.

LINN, R.L., BAKER, E.L. & DUNBAR, S.B. (1991) Complex, performance-based assessment: expectations and validation criteria, *Educational Researcher*, 20, pp. 15–21.

MADAUS, G.F. (1988) The influence of testing on the curriculum, in: L.F. TANNER (Ed.) *Critical Issues in Curriculum, Eighty-seventh Yearbook of the National Society for the Study of Education*, Part 1, pp. 83–121 (Chicago, IL, University of Chicago Press).

MESSICK, S. (1989) Validity, in: R. L. LINN (Ed.) *Educational Measurement*, 3rd edn, pp. 13–103 (New York, American Council on Education/Macmillan).

MESSICK, S. (1994) The interplay of evidence and consequences in the validation of performance assessments, *Educational Researcher*, 23, pp. 13–23.

MESSICK, S. (1995) Standards of validity and validity of standards in performance assessment, *Educational Measurement: issues and practice*, 14, pp. 5–8.

MOSS, P.A. (1992) Shifting conceptions of validity in educational assessment: implications for performance assessment, *Review of Educational Research*, 62, pp. 229–258.

MOSS, P.A. (1994) Can there be validity without reliability?, *Educational Researcher*, 23, pp. 5–12.

SHAVELSON, R.J., BAXTER, G.P. & GAO, X. (1993) Sampling variability of performance assessments, *Journal of Educational Measurement*, 30, pp. 215–232.

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

WIGGINS, G. (1993) Assessment, authenticity, context, and validity, *Phi Delta Kappan*, 75, pp. 200–214.