Validity in outcomes-based assessment

ROY KILLEN

ROY KILLEN *is an Associate Professor in the Faculty of Education and Arts at the University of Newcastle, Australia. He is also an Extraordinary Professor in the Faculty of Education at the University of Pretoria. He has worked extensively in South Africa over the past six years presenting seminars and workshops on assessment and outcomes-based education.*

Abstract

Because validity is universally accepted as one of the most important aspects of sound assessment practices, it is important that educators understand the concept and know how to use it as a quality control measure in their teaching. This paper outlines the evolving definition of the concept of validity, considers different types of evidence of validity and suggests how the concept might be applied in outcomes-based education. It argues that the common view that test items are valid if they measure what they are intended to measure, provides a necessary, but insufficient basis for considering validity. The paper develops the idea that test item per se can never be valid or invalid and explains why it is necessary to focus on evidence from which valid inferences can be made about learning.

Introduction

Most of those who write about assessment of learning claim that validity and reliability are the two most important characteristics of 'good' assessment items or tasks. It is therefore vital for teachers (including teachers in higher education) to be able to use these concepts appropriately to guide their assessment practices. Effective application of these concepts depends upon a deep understanding of their meaning and implications. Herein lies a problem: while the concept of reliability has remained reasonably static for the past half century, the concept of validity has evolved considerably. The 1950s view that a valid test was one that actually measured what you wanted it to measure has developed into the 1990s view that "validity is an integrated evaluative judgment 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" (Messick, 1989a, 13). Or, as Airasian (2001, 423) expresses it, validity is "the degree to which assessment information permits correct interpretations of the desired kind". The difference is not just one of detail; it is a significant change in emphasis, from validity being a property of a test item or assessment task to validity being a value judgement about inferences and actions made as a result of assessment – that is, a change in focus from the question "Is my test valid?" to the question "Am I making justifiable inferences and decisions on the basis of the assessment evidence I have gathered?" This represents an important shift in responsibility for validity, from the test constructor to the test user. To explore the importance of these changes, this paper will deconstruct and then reconstruct the concept of validity showing that it is a unitary but multifaceted concept that needs to be applied to the questions we ask, the overall assessment processes that we use, the interpretation of assessment results, and the inferences and decisions we make about student learning. The implications of this redefinition are described within the framework of outcomes-based education.

Why do we need a concept such as validity?

There are many reasons why we assess learners. Athanasou (1997) suggests that historically the principal purposes have been selection (for example, civil service entrance examinations), certification (for example, certification of people entering professions such as medicine) and classification (for example, intelligence tests). In more recent years the purposes of assessment have been broadened to include diagnosis (for example, determining why a learner has difficulty reading), grading (for example, grading students in a university subject), progression (for example, deciding whether individual students are ready to progress to the next year of school), programme evaluation (for example, assessing the merits of a particular instructional programme), and instructional improvement (for example, determining the effectiveness of particular instructional strategies). In all these applications of assessment, there are some important common threads. The first is a belief that there is something (intelligence, aptitude for a particular job, scholarly accomplishment, programme quality, and so on) that can be measured. The second is a belief that this factor can be measured in such a way that distinctions can be made between how much of it is possessed by different individuals, by different groups of learners or by different instructional programmes. Implicit in these beliefs is the idea that measurement can provide information that is accurate and worthwhile – information that can justifiably be used for making important decisions. The concepts of *reliability* and *validity* in educational measurement were developed in an attempt to describe just how worthwhile the information from a particular measurement might be and to suggest what might be done to increase the trustworthiness of measurement procedures.

In a strict measurement sense, reliability refers to "the degree to which test scores are free from errors of measurement" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1985, 8). There are several different ways of determining test reliability but for most practical purposes in teaching it is sufficient to say that "if a test is measuring something consistently, it is considered reliable" (Cunningham, 1998, 33). A very practical guide to enhancing reliability in assessment is provided by Herman, Aschbacher and Winters (1992) and this issue will not be considered further in this paper. However, reliability alone is not sufficient; a test could consistently measure something that was inappropriate or unrelated to the factor of interest. For a particular assessment task to provide useful information it must be reliable but it must also allow defensible judgements to be made about student learning. Hence the need for an additional criterion for judging the quality of our assessment tasks and practices – the criterion known as validity. The remainder of this paper will explore the concept of validity.

Deconstructing the definition of validity

Many books on assessment that are written for teachers define validity, rather loosely, as the extent to which "a test measures what it is meant to measure" (Hill, 1981, 22). Or, as Brady and Kennedy (2001, 55) put it, "when assessment tasks measure what teachers want them to measure, they are said to be valid tasks". There is some appeal in the simplicity of this definition, because it can serve as a useful starting point for discussion on test items or assessment tasks, particularly in outcomes-based education where the teacher can ask "is this item testing the outcome I want to test?" This simple view of validity as an inherent property of an item or test can be misleading and counterproductive. The reasons for this claim will be explored later in this paper, but first some of the historical developments in the concept of validity will be briefly reviewed.

There have been a number of significant stages in the evolution of the concept of validity. However it seems that the ideas emerging in each new stage have not always resulted in the majority of educators changing their assessment practices. In fact there is considerable evidence that many current assessment practices are still guided by vague conceptions of validity that are based on measurement theories from the 1950s (Messick, 1989a, 18). This has important consequences for the quality of assessment practices and for the assessment-based decisions that teachers make. Outdated conceptions of validity have the potential to mislead teachers, including those who are trying to implement outcomes-based education in South Africa.

Historically the conception of validity has developed in two important ways. First, there has been a development from the idea that a test is valid "for anything with which it correlates" (Guildford in Messick 1989a, 18) to the idea that there are a finite number of types of validity (Gronlund, 1982) to the idea that validity is a unitary concept (Messick, 1989a). Parallel with this change in conception of validity there has been a change in the reason why validity has been seen to be important. This change has essentially been a shift from prediction to explanation – from the idea that correlations between test scores and some criterion are important, to the idea that importance lies in the way in which test scores can be interpreted to provide information about some underlying construct. This change in focus emphasises that when these interpretations do not have a sound theoretical and empirical basis, their utility, relevance and value must be questioned.

Some textbooks written for teachers and teacher education students continue to emphasise the intermediate stage of this historical development – the stage that conceived validity as having three distinct types, namely content validity, predictive and concurrent criterion validity, and construct validity (see, for example, Brady & Kennedy, 2001). Furthermore, much of this literature gives prominence to the notion of content validity. Possible reasons for this situation will be explored briefly. Content validity is an indication of how relevant the content of a test or assessment task is, and how representative it is of the domain that is purported to be tested (Messick, 1989a, 17). Its determination is divorced from the responses that the test might invoke and from any inferences that might be made about student learning as a result of the test. It is essentially this concept of content validity that leads to claims such as "validity defines whether a test or item measures whatever it has to measure" (Van der Horst & McDonald, 2001, 185). The reasons why such a simplified view of validity appeals to teachers (and textbook writers) need to be considered. The appeal lies, foremost, in the simplicity; it is relatively easy to check that test items are addressing relevant content (or outcomes). It is also not too difficult to check that a test (or a series of assessment tasks) has adequate content coverage. Thus teachers can justify their assessment practices on the basis of uncomplicated and relatively easy-to-apply criteria. In so doing, teachers can (perhaps unintentionally) absolve themselves from considerable responsibility; they can feel content that their responsibility ends with the creation or selection of a test that has domain relevance and representativeness. The problem is that such an approach ignores the importance of the conditions under which the test is administered, the effect that learners' characteristics will have on their responses, and the responsibility that teachers have to interpret the test results in defensible ways. This simplified approach to validity also overlooks the fact that determining what a test measures requires more than considering just content relevance and representativeness.

One way to address this shortcoming is also to consider criterion-related validity. Historically, the criterion-related validity of a test was determined by comparing the test scores with one or more external variables (called criteria) that were considered to provide a direct measure of the behaviour or characteristic in question (Messick, 1989a, 17). The comparisons were usually made by calculating correlations or regressions. Criterion-related validity was a useful concept

in situations where a standardised test was used repeatedly with different groups of subjects after its correlation with a direct measure had been established. One reason why most teachers give little attention to this form of validity is that appropriate external criteria are generally not available. For concurrent criterion-related validity to be determined, teachers would need access to external criteria that provided direct measures of the learning they were trying to assess. If such direct measures were available, the sensible thing would be for teachers to use those direct measures, rather than some indirect measure that they developed themselves. This is generally not the situation in classroom assessment where it is more likely that individual teachers are developing tests that have not been correlated with any external, direct measures of whatever they are trying to assess.

A similar situation exists when we consider predictive criterion-related validity. This concept was used historically to describe the correlation between a test score and some criterion measurement made in the future (Messick, 1989a, 17). Again, this was a useful concept when the external measure was established as a direct measure of the quality of interest and if the earlier test was standardised and used repeatedly with different groups of learners. However, this is not the situation with most teacher-developed tests. By definition the predictive validity of a test cannot be determined until the subjects (or a comparable group of subjects) have been tested on the (future) criterion test. For normal testing purposes in classrooms, this means that the predictive validity of teacher-developed tests cannot be determined in a timeframe that makes the exercise worthwhile. Thus, criterion-related validity that is concerned just with specific test-criterion correlation is not a particularly useful concept for classroom teachers.

This then leads to a consideration of the third historical type of validity – construct validity. Constructs are theoretical conceptual frameworks for describing human characteristics, behaviours or groups of abilities. For example, reasoning ability, self-esteem and communications style are hypothetical constructs. Construct validity is essentially concerned with investigating the meaning of test scores. It is based on the idea that a score on a well-constructed test can be taken as one (of possibly many) indicators of the construct of interest. For example, a test might provide a strong indication of a learner's mathematical reasoning ability (the construct of interest), but it would be just one of many such indicators. Likewise, the test might be a strong indicator of other constructs (such as reading ability). Historically, test construct validity was evaluated by examining patterns of relationships among item scores and patterns of relationships between test scores and other external measures. This was later extended to include studies of performance differences over time, across groups and settings and in response to experimental treatments. Investigations such as these led to the suggestion that construct validity (making meaning of the measurement) actually subsumed considerations of content relevance and representativeness, and criterion-relatedness (Messick, 1980).

The practice of thinking about different aspects of validity as separate entities was common until the mid-1980s and has persisted in many modern texts. For example, Brady and Kennedy (2001) distinguish between five separate aspects of validity (content validity, construct validity, consequential validity, concurrent validity and predictive validity). The problem here is not that validity is being considered as multidimensional but that the gestalt is being lost in the detail. It is more productive to think of validity as a unitary concept, but not as a simplistic one. This point will be illustrated through an example.

Recently, the author was working with three high school mathematics teachers to help them improve their teaching of basic algebra. Initial discussions focused on a recent test that the teachers had given their students. The teachers were satisfied that the test was valid because for them it measured what they wanted it to measure – knowledge of basic algebra and ability to

solve simple algebraic equations (such as finding the value of x in the expression $3x + 5 = 20$.) As a result, the teachers felt justified in concluding that students who did well in the test had a 'good' knowledge of basic algebra and students who scored low marks in the test had a 'poor' knowledge of basic algebra. The teachers' notion of validity did not extend beyond the idea that the questions tested what they wanted them to test.

When the teachers were prompted to consider validity from the perspective provided by Brady and Kennedy (2001), they were satisfied that the test had "content validity" (the questions were linked to the curriculum content), "construct validity" (they believed that the test was indicative of the students' broader understanding of algebra), "concurrent validity" (students' results in this test were very similar to their results on other algebra assessment tasks completed at about the same time), and "predictive validity" (past experience indicated that students' performance in this test was indicative of the results they might obtain on later algebra tests).

Following discussions with the teachers, the author interviewed a group of students who had scored low marks in the algebra test. This revealed several important pieces of information, including the following: some students were incapable of performing basic arithmetic calculations (for example, they could not multiply 2.5 by 8); and, all students in the group had very limited understanding of fractions and were unable to perform simple calculations involving fractions (such as 6 x ¾). These two factors were responsible for approximately 50% of the errors made by these students in the test.

This deeper investigation of the students' results suggested that no matter how well the test assessed the algebraic knowledge of *some* students in the class, the measures of validity described above were inappropriate ways of considering the results of the students who did not perform well in the test. It mattered little that the test was aligned with the curriculum content (the students were incapable of demonstrating their understanding of that content because they did not have the necessary arithmetic ability). The test did not necessarily indicate the students' understanding of algebraic concepts because their arithmetic skills limited their ability to demonstrate that understanding. It was highly likely that their results in this test would be similar to their results in other tests that relied on the same prior knowledge. The apparent concurrent validity and predictive validity were not due to the students' knowledge of algebra, but to the students' underlying lack of prior knowledge of arithmetic.

The above analysis does not mean that the test was of no value, but it does highlight an extremely important issue. No matter how we try to distinguish between different types of validity (construct validity, content validity, and so on) it is simply inappropriate to say that a test item or assessment task is valid in some absolute sense. Rather, we should think of validity in terms of the definition provided by the American Educational Research Association, American Psychological Association and the National Council on Measurement in Education (1985, 94) namely that validity is a unitary concept that refers to the "degree to which a certain inference from a test is appropriate and meaningful". From this perspective, assessment tasks themselves can never be valid or invalid: it is the assessment-based inferences we make that are valid or invalid. This distinction is important because it addresses one of the fundamental challenges of outcomes-based education – the challenge of developing assessment instruments that allow teachers to draw valid inferences about the extent to which learners have achieved curriculum outcomes. In the example above, the inference that students who scored high marks in the test had a good understanding of basic algebra *may* have been valid, but the inference that students who scored low marks in the test had a poor knowledge of algebra may have been invalid because the particular test items just happened to involve arithmetic operations that they could not perform. As Messick (1989a, 42) points out, all that can be claimed about low scorers

is that they did not perform the task successfully and "there is no basis in test performance *per se* for interpreting low scores as reflective of incompetence". Before such an inference can be made, it is necessary to discount other plausible explanations for the low scores. The factors that can lead to irrelevant variance in test performance might include anxiety, fatigue, inattention, low motivation, limited language proficiency or (as in the above example) lack of prior knowledge. Trying to assemble evidence that will explain variations in test scores is, Messick (1989a, 42) argues, "the hallmark of construct validation".

The message here is simple, but it is frequently overlooked: You cannot validate a *test*; you can only validate the *inferences* that are drawn from students' results in the test. Although "evidence can be accumulated in many different ways, validity always refers to the degree to which the evidence supports the inferences that are made" (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1985, 1). Or, as Popham (1990, 94) argues, test items or other assessment tools simply yield data (typically test scores) and "the *interpretation* of those test scores is the operation which may or may not be valid". This is the point that teachers often overlook when thinking about validity – they focus on the questions they have asked, not on the inferences they have drawn. The mathematics teachers in the above example had done this – they had satisfied themselves that the test was valid (because they thought it measured what they wanted it to measure), and they ignored the invalidity of the conclusions they had reached about the low-scoring students. The teachers did not realise that the test questions did not allow them to draw valid inferences about the knowledge of some learners, even though the test did provide evidence from which valid inferences could be drawn about the knowledge of other learners.

One of the major consequences of a 're-definition' of validity (so that it focuses on the inferences rather than the test items) is that it places the responsibility for validity squarely with the educator who makes the inferences. No longer can a teacher claim to be using a valid assessment task simply because it is clearly linked to the curriculum content, or because someone else has used the test and decided that it is valid, or because it produces results similar to those obtained from other assessment tasks. Instead the teacher must question the validity of the inferences *they* are making as a result of having used the assessment task. The teacher's challenge is no longer to develop valid assessment tasks, but to develop assessment tasks that will generate evidence from which valid inferences can be drawn about the learning of all students. The distinction between these two views of validity is not a trivial one – high quality assessment tasks are necessary, but not sufficient, for making valid inferences about student learning. Some of the ways in which validity of inferences can be tested will now be examined.

Evidence of validity

From the above discussion it is clear that we should be seeking evidence that assessment-based inferences and decisions are valid, rather than seeking evidence that assessment tasks *per se* are valid. From this perspective, "evidence of validity" can be sought in several ways. In each case we need to consider both the assessment tasks and the processes used to draw inferences from students' attempts at these tasks. The 1985 *Standards of the American Educational Research Association, American Psychological Association and the National Council on Measurement* provide a useful starting point for this investigation. They suggest that it is necessary to consider *content-related evidence of validity*, *criterion-related evidence of validity* and *construct-related evidence of* validity – that is, evidence that the inferences have a sound basis in terms of the curriculum content that is being assessed, the criterion measure and the broad constructs that are being assessed. The distinction between considering types of validity and considering types of validity *evidence* may seem a pedantic one, but it was an important

historical step in the study of validity (Messick, 1989a, 20). These concepts will be explained briefly and then expanded to provide a view of validity that is appropriate in outcomes-based education.

There are two aspects of *content-related evidence of validity*. First, it is necessary to consider whether or not the assessment task is capable of producing content-related evidence from which valid inferences can be drawn. Second, it is necessary to consider the degree to which this evidence is actually used to draw valid inferences. One of the most useful ways to determine whether or not a test is capable of providing a valid basis for making content-related inferences about student learning is to have a suitable panel of judges evaluate the test. The judges should focus on two things: (1) whether or not each item in the test is related directly to the curriculum content that is to be tested; and (2) the extent to which the total assessment task is capable of measuring an appropriately representative sample of the important components of the curriculum content. Popham (1990) refers to these two issues as *item relevance* and *content coverage*. For practical purposes in schools, the 'panel of experts' could be a group of teachers from a relevant subject area.

Item relevance and content coverage describe the *potential* of a test to provide evidence from which valid inferences can be drawn. If a test item is not related to the curriculum content then it cannot produce useful evidence of student learning. If the total assessment task does not test a suitably representative sample of important curriculum content then there will be insufficient evidence from which to draw valid inferences. However, although item relevance and content coverage are necessary, they are not sufficient to guarantee that valid inferences are drawn. It is possible to have test items that are relevant and representative, but inappropriate inferences drawn from students' answers to these questions. For example, a question that asks students to explain the principles of outcomes-based education might be quite appropriate in a course on curriculum development. However, if students' answers to the questions are not aligned with the teacher's views on OBE, the teacher may draw inappropriate inferences about how well the students understand OBE (even if their understanding is deeper than the teacher's). Therefore, in addition to considering the relevance and content coverage of the assessment items, it is necessary to consider the extent to which the inferences drawn by the teacher can be justified.

The concept of validity changes considerably when we focus attention on the inferences or conclusions drawn by the teacher rather than just on the test items. Validity is no longer an inherent property that can be determined coldly and objectively by analysing test items. It becomes a subjective judgement on the actions (and, by implication, the expertise) of the teacher. It will now be necessary to ask questions such as: What evidence is there that the inferences drawn by the teacher were based solely on the evidence produced by the test? Were the inferences clearly related to important content that was within the bounds of the curriculum? On what basis can each inference be justified? Has the teacher making the inferences taken account of all relevant information? This shift in emphasis will require teachers to rethink the way they evaluate test items, but it is still not a sufficient description of validity.

There are dangers in assuming that valid inferences are being made from test results without testing the legitimacy of these assumptions. A brief example will be used to illustrate this point. The author recently reviewed a series of examination papers for a university subject. The multiple-choice questions used in these papers were taken from an item bank that had been developed over several years. Before each examination, a group of lecturers collaborated to select test items that they considered to cover a representative sample of the subject content. After each examination, the results had been analysed using standard procedures for calculating the difficulty and discrimination indices of each item and items that were identified as

problematic were modified or removed from the item bank. Each year, the results in these examinations were used to determine whether students had passed or failed the subject. Implicit in these decisions were assumptions that the examinations were an appropriate means of testing the outcomes of the subject and that the inferences about students' achievement of the outcomes were valid.

A closer analysis of the examination items and the students' results revealed a number of problems, including the following. All the questions had "face validity" (Cunningham, 1998, 42) because each one appeared to measure what it was supposed to measure. Indeed, the test constructors had gone to considerable trouble to link each item with a specific section of the subject textbook. However, there was little content-related evidence of the validity of decisions concerning students' achievement of outcomes because (a) the majority of items were not testing *important* aspects of the content, (b) because the majority of questions were testing low levels of cognition there was no evidence that students were achieving the subject outcomes that required higher-order thinking, and (c) because the complex wording of some questions biased them against students from non-English speaking backgrounds. These difficulties meant that although the test items appeared to be valid (from a 1970s perspective of validity), the inferences being drawn from the test results were inappropriate. Difficulties such as these can be minimised if assessors repeatedly remind themselves that the purpose of assessment is to gather evidence from which inferences about student learning can be drawn, and repeatedly seek content-related evidence that these inferences are valid. Instead of asking "Is this a valid test item?" assessors should be asking questions such as "Will this test item provide me with evidence from which I can draw valid inferences about student learning?" and "Will it provide equally useful evidence from all learners, or only from some learners?"

The analysis of these questions highlighted a problem that is discussed by Herman *et al.* (1992, 102) who point out that "while the task on the surface may appear to assess desired outcomes, until you see the actual student responses, you cannot be completely clear about what you are measuring". In this case, it was only when the errors students were making were examined closely that the language problems became clear. The evaluation exercise also highlighted the dangers of basing assessment on a single type of question. It gave some credence to the claim by Herman *et al.* (1992, 102) that "the only way to know whether the assessment really assesses your intended goals is to gather evidence corroborating the test score interpretation".

The second type of evidence of validity is *criterion-related evidence*. There are many situations in which educators want to measure one thing and determine whether or not it is systematically related to something else. For example, they might be interested to know whether students' results in an assessment task completed at home are a good indication of their future examination performance (the *criterion* measure). Therefore they are interested in the criterion-related evidence that inferences they make about the relationship between the assessment task and the examination results are valid. Again, it is not a simple process of judging the validity of the assessment task, the criterion measure and the relationship between the two. Most significantly, it is necessary to examine the evidence that the information on the tasks and their relationship has been used in appropriate ways to draw defensible conclusions relating to student learning.

In addition to considering content-related evidence and criterion-related evidence of the validity of the inferences we make about student learning, it is also necessary to consider *construct-related evidence of validity*. This involves seeking evidence that the assessment task is actually providing a trustworthy measurement of the underlying construct in which we are interested. If this can be established, then we have construct-related evidence that the inferences we make,

based on the test, have a *possibility* of being valid. However, it is still necessary to consider whether or not the inferences actually *are* valid.

There is a strong argument that all validity should be defined as a form of construct-related validity (Messick, 1989a; Cunningham, 1998). Messick (1989b, 8) argues that "fundamentally, all validation is construct validation, in the sense that all validity evidence contributes to (or undercuts) the empirical grounding or trustworthiness of the score interpretation". From this perspective, every assessment task should be considered as a measurement of a construct or a series of constructs and assessors should attempt to integrate all the evidence that bears on the interpretation of the learners' performance in the task. The inferences drawn from the results of assessment (and the actions taken as a result of those inferences) should be meaningful and trustworthy. For example, an assessment task might be designed to measure "decision-making ability". This construct can be defined in terms of specific knowledge, specific skills and certain attitudinal characteristics (Wood *et al*., 2001). To be satisfied that the task provided evidence from which valid inferences could be drawn about learners' decision-making ability, it would be necessary to consider the content relevance and coverage of the task, the evidence that task performance was indicative of performance on other reliable measures of decision-making ability, and evidence of the degree to which the defining elements of the construct actually influenced learners' performance on the task. This evidence would help to determine the meaningfulness of the interpretation of the test results. Messick (1989a, 13) argues that "construct validity is the integrating force that unifies validity issues into a unitary concept". This idea will now be explored in relation to OBE.

Redefining validity in outcomes-based education

The ideas explored earlier in this article provide a sound basis for considering how validity could be conceptualised in an outcomes-based education system. The general principle of seeking content-related evidence, criterion-related evidence and construct-related evidence to support claims that the inferences drawn from assessment results are valid is still appropriate. However, in outcomes-based education there are some special aspects of assessment that need to be addressed before we can claim to be making valid inferences about student learning.

Although it is a useful starting point, the definition of content-related evidence of validity in the 1985 Standards of the American Educational Research Association, American Psychological Association and the National Council on Measurement in Education gives the impression that learning should be focused primarily on the accumulation of knowledge (content). In outcomes-based education, it would be more appropriate to require *outcome-related evidence of validity* – that is, evidence that we are drawing valid inferences about achievement of outcomes, rather than about learning of content. Following Popham's (1990) suggestions about quantifying content-related evidence of validity, we could think of outcome-related evidence of validity as having two components, relevance and outcome coverage. That is, we could determine whether or not each assessment item required that learners demonstrate one or more relevant outcomes and we could determine the extent to which the total assessment task or programme addressed an appropriately comprehensive sample of outcomes.

Establishing that an assessment task has an appropriate level of relevance and outcome coverage is important, but it is not sufficient. Alignment between outcomes and assessment is the foundation for drawing valid inferences about student learning, but it does not guarantee that valid inferences will be drawn. The second aspect of *outcomes-related evidence of validity* is evidence that the teacher has drawn valid inferences about learners' abilities to demonstrate the learning that is either explicitly or implicitly described by the outcomes. This requires

evidence that the teacher has (a) taken into account the characteristics of the learners (e.g., their language proficiency), (b) interpreted the students' responses appropriately, (c) recognised the full range of abilities that the task is assessing, and (d) not extrapolated from the results to draw inferences about learning that has not been tested.

One of the sources of invalidity in inferences about learner achievement in OBE is the belief that it is possible to make clear-cut decisions about whether or not a learner has achieved a particular outcome. In outcomes-based education it is often said that the principal reason for assessment is so that we will know whether or not learners have achieved the outcomes we wanted them to achieve. This view of assessment is unfortunate because it can mislead teachers into thinking that it is possible to make clear distinctions between those learners who have achieved certain outcomes and those who have not. This view is often the result of a narrow interpretation of Spady's (1994) definition of OBE. In Spady's words:

> Outcome-Based Education means clearly focusing and organizing everything in an educational system around what is essential for all students to be able to do successfully at the end of their learning experiences. This means starting with a clear picture of what is important for students to be able to do, then organizing the curriculum, instruction, and assessment to make sure this learning ultimately happens" (Spady, 1994, 1).

According to Killen (2002), this definition is based on the implicit assumptions that someone can determine what things are "essential for all students to be able to do", and that it is possible to achieve these things through an appropriate organisation of the education system and through appropriate classroom practices. Most significantly, it also implies that it is possible to determine, in some objective way, whether or not learners have achieved whatever outcomes were deemed to be important. This view, coupled with Spady's (1994) recommendation that outcomes should always contain an 'action verb', often leads to the assumption that assessment decisions can be reduced to placing learners into categories such as "achieved/not achieved" or "exceeded/satisfied/ partially satisfied/not satisfied" (Department of Education, 2002) for each outcome. For all but the simplest of outcomes, this type of categorisation is almost impossible to justify and it results in invalid judgements and inferences.

To overcome this problem, it is necessary to think of assessment as the process of determining *how well* learners are able to demonstrate what they have learned, rather than trying to determine in some categorical sense which learners have or have not learned. If this approach is coupled with acceptance of the idea that outcomes can legitimately be expressed in terms of "understanding" (Van Niekerk & Killen, 2000) and not just in the behavioural terms recommended by Spady (1994), the foundation is set for valid inferences to be drawn about students' learning. It changes our focus from asking "How many questions can the learner answer?" or "Which skills can the learner demonstrate?" to "How well does the student answer questions?" and "How expertly can the learner demonstrate particular skills?" Understanding (rather than memorisation), creativity (rather than reproduction), diversity (rather than conformity), initiative (rather than compliance) and challenge (rather than blind acceptance) become the yardsticks by which we try to measure, describe and report student learning. When this approach is taken, it is not appropriate to think of an assessment task *per se* as valid. The tasks will deliberately be designed to evoke responses that are indicative of each learner's level of understanding; they will not do this equally well for all learners. Therefore, educators will need to focus on the extent to which a particular assessment task enables individual learners to demonstrate their understanding and, most importantly, will have to consider the extent to which the available evidence can support legitimate inferences about the learning of each student. The implications of this new focus are quite significant.

Before we can draw appropriate inferences about student learning from assessment tasks that are designed to elicit a range of responses from learners, we have to clarify what it means to learn and understand things in the subject area in which the assessment will be grounded. Put simply, you cannot assess what you cannot define. Furthermore, you cannot draw valid inferences about student learning that has not been assessed n appropriate ways. This presents a particular challenge in higher education because much of what we want students to learn is difficult to define in precise terms – the nature of the knowledge with which we are dealing often means that outcomes can be written only in terms of quite complex, ill-defined concepts. It is, for example, much more difficult to define what we mean by "understanding the principles of curriculum design" than it is to define what we mean by "being able to perform simple arithmetic operations". When we attempt to define what we want students to learn, we may decide that understanding is the capacity to use explanatory concepts creatively, or the capacity to think logically, or the capacity to tackle new problems, or the ability to re-interpret objective knowledge – to mention just a few possibilities. Quite clearly, a particular educator's belief about what it means to "understand" will influence the way that person tries to help learners to understand, how they attempt to assess their learners' understanding, and the inferences they draw from students' attempts to demonstrate their understanding. Thus, the validity of our interpretation of assessment evidence cannot be divorced from our conceptions of learning.

Consider, for example, the implications for assessment of the view expressed by Kissack (1995, 260) that studies in higher education, particularly in the humanities or social sciences, should not be concerned with the memorisation and reproduction of voluminous amounts of information, but rather with "an appreciation of the nature of the 'object of knowledge' ... and an ability to construct and defend the particular arguments which constitute the individual's interpretive perspective". This view will lead to the formulation of outcomes such as "Learners will be able to develop and defend a personal interpretation of the concept of postmodernism in literature". Quite clearly, an outcome such as this calls for assessment tasks that will produce qualitative rather than categorical evidence of learner achievement. Evaluation of this evidence of learning will require high levels of interpretation and subjective judgement and it is the validity of these interpretations and judgements that will have to be defended, not the validity of the questions that prompt the generation of the evidence.

When we progress beyond trivial behavioural outcomes, it becomes obvious that there is not a clear cut-off between achieving/not achieving the outcome. We are no longer faced with the problem of inferring whether or not a learner has achieved a particular outcome. Instead, we have the challenge of making valid inferences about *how well* each outcome has been achieved. Essentially this comes down to a need to be able to describe (in words rather than numbers) the difference between low-quality and high-quality achievement of the outcome. Until this can be done, it is not possible to devise an assessment task that will allow learners to demonstrate the quality of their learning and, therefore it is not possible to draw valid inferences about their learning.

Defining the level or standard of achievement of a non-trivial outcome is a complex process that can be approached in several different ways. One useful approach is to apply the *Taxonomy of Learning, Teaching and Assessing* (Anderson & Krathwohl, 2001). They suggest a six-level hierarchy of cognition (remember, understand, apply, analyse, evaluate and create) and four types of knowledge (factual, conceptual, procedural and metacognitive) that combine to form a two-dimensional grid onto which outcomes can be mapped. This helps to clarify the outcome and to simplify the process of aligning outcomes, teaching strategies and assessment. It emphasises, for example, that an outcome requiring application of procedural knowledge must be assessed quite differently from an outcome requiring analysis of factual knowledge. Aligning

the assessment procedures with the required outcomes is essential if valid inferences are to be drawn about the responses of learners. However, this is not sufficient, it is also necessary to consider the possible levels of achievement of the outcome. The standards-referenced assessment framework described by Killen (2000) and the SOLO taxonomy (Biggs & Collis, 1982) are both useful for this purpose. Either method will provide a detailed description of multiple levels of achievement of the same outcome, thus helping to define the focus for assessment tasks and the focus for the inferences we make about student learning.

When we seek evidence concerning the correspondence between the intended cognitive demands of the assessment task and the cognitive activity that the task actually invokes in the respondent we focus on what Glaser and Baxter (1997, 3) refer to as "cognitive validity". This view of validity also seeks evidence of the correspondence between the quality of cognitive activity invoked by the task and the performance scores given to respondents. The critical issue here is that different forms of an assessment task, each with appropriate outcome relevance, can require respondents to use quite different cognitive processes, leading to possibly inappropriate conclusions about their level of understanding. For example, two tasks with equivalent outcome relevance might provide students with quite different levels of directedness (or freedom to respond in ways that they devised for themselves). It may even be that the structure of the task allows learners to respond "correctly" for reasons that are unrelated to their understanding of the construct being tested – this source of invalidity is referred to by Messick (1995) as *construct-irrelevant variance*. To minimise this source of invalidity, different versions of an assessment task should be compared in terms of (a) intended task demands, (b) inferred cognitive activities that underlie the task, and (c) scores obtained (Glaser & Baxter, 1997).

Just as there cannot be a yes/no determination about whether a learner has achieved an outcome that is not trivial, there cannot be a yes/no determination of whether an inference drawn from assessment-generated evidence is valid. Rather, we need to consider the factors that will enhance the likelihood that valid inferences will be drawn and try to minimise those factors that are likely to diminish the validity of our inferences. The starting point must be a clear definition of the outcomes we want learners to achieve and a clear conception of the theoretical constructs that define and structure the area of study. We then need to consider the theoretical and practical justification for inferring that the assessment process has the potential to provide trustworthy evidence about student learning. Finally, we need to consider the extent to which the inferences drawn about student learning are justifiable.

Conclusion

Outcomes-based education requires teachers to focus on helping learners to apply the things they learn rather than simply to accumulate knowledge. This makes it necessary to use forms of assessment that require more than the reproduction of 'content'. Consequently, the historical view that valid test items are those that test what they were intended to test is too narrow and can lull teachers into a false sense of security about the quality of their assessment practices. Rather than thinking of an assessment task as valid simply because it appears to assess certain content or a particular outcome it is also necessary to focus on the validity of the judgements teachers make and the inferences they draw from the assessment-generated evidence. This process should start with a consideration of the appropriateness of the learning outcomes and be followed by a consideration of the extent to which the learning opportunities made it possible for students to achieve these outcomes to high levels of proficiency. Next, the assessment tasks should be evaluated in terms of their outcome relevance and coverage and their potential to provide useful evidence about the constructs that they are designed to measure. Finally, the evidence produced from these tasks should be interpreted in defensible ways. It is only when all

these elements are in place and aligned that the inferences drawn about student learning have the potential to be valid.

The important challenge for teachers is not to construct valid test items *per se*, but to construct test items, administer tests, mark and interpret results in ways that will allow valid inferences to be made about student learning. If this challenge is not addressed, teachers will be ignoring the single most important characteristic of assessment – its ability to help them make appropriate instructional decisions.

References

Airasian, PW. 2001. *Classroom assessment: Concepts and applications*. 4th ed. Boston: McGraw Hill.

Anderson, L & Krathwohl, D. 2001. *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

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

Athanasou, JA. 1997. *Introduction to educational testing*. Katoomba, Australia: Social Science Press.

Biggs, JB & Collis, KF. 1982. *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. New York: Academic Press.

Brady, L & Kennedy, K. 2001. *Celebrating student achievement: Assessment and reporting*. Sydney: Prentice Hall.

Cunningham, G. 1998. *Assessment in the classroom*. London: The Falmer Press.
Department of Education. 2002. *Revised National Curriculum Statement R-9 (Schools)*. Pretoria: Department of Education.

Glaser, R & Baxter GP, 1997. *Improving the theory and practice of achievement testing*. Paper presented at the BOTA meeting, National Academy of Science/National Research Council, Washington, DC, February.

Gronlund, N. 1982. *Constructing achievement tests.* 3rd ed. Engelwood Cliffs, NJ: Prentice-Hall.

Herman, JL, Aschbacher, PR & Winters, L. 1992. *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Hill, J. 1981. *Measurement and evaluation in the classroom*. Columbus, OH: Merrill.

Killen, R. 2000. *Standards-referenced assessment: Linking outcomes, assessment and reporting*. Keynote address at the Annual Conference of the Association for the Study of Evaluation in Education in Southern Africa, Port Elizabeth, South Africa, 26-29 September.

Killen, R. 2002. Outcomes-based education: Principles and possibilities. *Interpretations*, **35**(1), 1-18.

Kissack, M. 1995. Hermeneutics in education: Reflections for teachers of the humanities. In *Metatheories in philosophy of education*, Higgs, P. (ed.). Johannesburg: Heinemann.

Messick, S. 1980. Test validity and the ethics of assessment. *American Psychologist*, **35**, 1012-1027.

Messick, S. 1989a. Validity. In *Educational measurement*, 3rd ed., Linn, RL (ed.). New York: Macmillan.

Messick, S. 1989b. Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, **18**(2), 5-11.

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

Popham, W. 1990. *Modern educational measurement: A practitioner's perspective*. 2nd ed. Englewood Cliffs: Prentice Hall.

Spady, W. 1994. *Outcome-based education: Critical issues and answers*. Arlington, VA: American Association of School Administrators.

Van der Horst, H & McDonald, R. 2001. *Outcomes-based education: Theory and practice*. Pretoria: Van der Horst & McDonald.

Van Niekerk, L & Killen R. 2000. Recontextualising outcomes-based education for teacher education. *South African Journal of Higher Education,* **14**(3), 90-100.

Wood, J, Wallace, J, Zeffane, RM, Schermerhorn, JR, Hunt JG & Osborn, RN. 2001. *Organisational behaviour: A global perspective.* 2nd ed. Milton, QLD: John Wiley.