20 Assessment of Knowledge with Written Test Forms

STEVEN M. DOWNING
University of Illinois at Chicago

SUMMARY

Written or computer-based assessment is the most appropriate modality to test cognitive knowledge. Two major classes of written assessment are discussed in this chapter: selected-response and constructed-response formats. Multiple-choice (MCQ), matching and extended matching, true-false, multiple-true false and alternate-choice items are examples of the selected-response format. Long and short essays, very short essays, modified essay questions and written simulations, such as computer simulations, are examples of constructed-response formats. These assessment forms are critically evaluated in terms of their strengths, limitations, appropriateness, efficiency, research basis, psychometric properties, and validity evidence.

INTRODUCTION

This chapter discusses written (and computer-administered) cognitive test forms or formats intended to assess learner achievement. Cognitive assessments are the most common and, arguably, the most important types of evaluation used in medical education settings of all kinds throughout the world. Performance examinations – tests of actual or simulated psychomotor skills or affective behavior – are discussed in other chapters in this section.

Two major types of written cognitive assessments are discussed – selected-response and constructed-response item formats. These names are accurately descriptive of the forms used. Selected-response formats require the examinee to *select* answers from a listing of possible responses; the multiple-choice item form is the most typical example of selected-response, but formats such as the matching and true-false forms are also representative of this format. Constructed-response formats challenge examinees to *produce* responses to more open-ended questions or stimuli. Examples of constructed-response formats include essays – long and short

International Handbook of Research in Medical Education, 647–672.
G.R. Norman, C.P.M. Van der Vleuten, D.I. Newble (eds.)

essays and very short-answer essays, modified essay questions and the simulation formats.

The chapter begins with a general discussion of important concerns in testing cognitive knowledge, the generally accepted characteristics of adequate achievement measurement, and the arguments for and against selected-response and constructed-response items. This same structure is carried into a discussion of each particular item format with examples of most item forms.

TESTING COGNITIVE KNOWLEDGE

In order to discuss testing cognitive achievement using written test forms, it is necessary to first operationally define cognitive knowledge in this context. *Cognitive knowledge* includes all learning or achievement that is associated with some theory of mental ability or function. Examples are theories modeled by taxonomies such as described by Bloom (Bloom, Engelhart, Furst, Hill, & Kratwohl, 1956) or information processing models such as discussed by Anderson and Bower (1973) or by cognitive psychology models (e.g., Mislevy, 1993; Snow & Lohman, 1989; Dibello, Roussos, & Stout, 1993). Operationally, cognitive knowledge assessment encompasses all testing intended to measure examinee *mental* learning that is not in the psychomotor or affective domain. Thus, important inferences from test scores are made to some domain of cognitive ability or achievement, rather than some curricular or instructional domain (Millman & Greene, 1989) or some psychomotor (performance) or affective domain.

Tests of cognitive knowledge

Cognitive knowledge is best assessed using written test forms. This bold statement is supported by some eighty years of educational measurement research on objectively scored written examinations. Since almost all cognitive knowledge is verbally mediated, it is of no surprise that verbal (written) assessments are the primary and most desirable forms used to test achievement in this domain. Written test forms, on the other hand, are almost useless as measures of psychomotor or affective skills, abilities, and achievements (with some reservations).

Desirable characteristics of cognitive tests

Cognitive assessments all share a set of desirable characteristics which most educational measurement professionals would agree should be present for adequate assessment of achievement. These desirable characteristics are necessary but not sufficient conditions for good, trustworthy tests of achievement. They are minimum

requirements for adequate measurement, but many other attributes may be needed for certain specific types of tests to ensure appropriate inferences from test scores.

Objectivity

Objectivity refers to the characteristic of agreement among content experts that the correct or keyed response is indeed correct or that a constructed answer meets some minimum requirement for credit. Objectivity generally implies that examinee responses can be machine or computer scored since there is little or no subjectivity or judgment involved in deciding whether a particular examinee's answer or response is right or wrong. In its most basic form, objectively scored item formats can be reduced to a set of ones and zeros for each correct and incorrect examinee answer.

Selected-response item formats are most readily and easily objectively keyed. It is much easier for experts to agree on the single-best answer or the most-correct-response out of a set of carefully crafted and edited possible answer options than it is to agree upon a scoring rubric or the minimum essential statements required for a more open-ended constructed-response item.

Constructed-response formats can be objectively scored, but the task is generally much more difficult than with the selected-response formats. Clearly some types of achievement assessment, such as the evaluation of writing skill, require expending the additional effort needed to maximize the objectivity of measurement with constructed-response items.

Measurement properties

All achievement measurements must have certain properties in order to produce test scores that are useful, trustworthy and provide legitimate and accurate inferences. Some of these important properties are: score scales which have adequate validity evidence, are reproducible, and correspond roughly one-to-one with levels of cognitive knowledge attained by examinees.

Some basic definitions of terms may useful:

Test scores

A test score is the quantitative value (the number) or other symbol (a letter) produced by counting the number of correct responses to test questions. Test Score Scales are the set of numbers produced by a group of examinee test scores. Some examples of score scales are: raw score scales which are the simple sum of the number of correctly answered items; transformed scales are raw scores converted into some other metric, such as the percent-correct scale or a standard score scale which is a linear transformation of the raw scores into a scale with the mean and standard deviation of scores set to some arbitrary value.

Validity evidence

The contemporary view of test validity suggests that tests are not valid or invalid, but rather it is the interpretation of test scores and the inferences drawn from those scores that are more or less valid. Evidence is collected, sometimes from many different sources and over a long period of time, to support or refute the claim for the reasonableness of the conclusions drawn from test data or the inferences about examinee achievement in the domain measured or the construct sampled (Messick, 1989; Kane, 1992).

There are direct sources of validity evidence for many achievement tests. The content tested by the questions selected for the examination is the most critical source of validity evidence for most achievement tests. How the content is selected, and in what proportions to the population or domain of interest, are important sources of validity evidence. The qualifications and expertise of the item writers and test constructors may be another source of validity evidence for such examinations. More indirect sources of validity evidence for achievement examinations are statistical relationships of test scores with other measures (such as other tests of similar content) or ratings of clinical performance by independent judges or class rankings.

In general, the more evidence one has to support the intended inferences from test scores (validity evidence), the better. It is not possible to have too much validity evidence to support the specific claims for valid inferences. Also, validity evidence must be constantly updated, as examinees, educational programs, and test questions change and evolve. Validity evidence may have a fairly short shelf life and the more important the consequences of the test scores, the more important it is to update validity evidence.

Reproducibility

Reproducibility refers to the test score characteristic of score and pass-fail decision reliability. Just as with any scientific experiment, the ability to independently reproduce test results and the pass or fail decisions that follow from these scores is critical to the trustworthiness of the test results. Theoretically, a second administration of the same examination to the same examinees (assuming that they have learned or forgotten nothing) should produce about the same test scores, if the test is well constructed. On the other hand, if test results are markedly different on the second administration (we are unable to reproduce the results), one questions the adequacy of the test to measure the content or construct of interest.

Many methods are available to estimate the reproducibility of test scores and the pass-fail decisions resulting from the test scores. For most achievement examinations, an internal-consistency reliability coefficient such as the Kuder-Richardson Formula 20 (Kuder & Richardson, 1937) or the more general formula, Cronbach's Alpha (Cronbach, 1951), is an adequate estimate of score reproducibility. Generalizability Theory (Brennan, 1983) permits parsing more

sources of measurement error and may be an appropriate estimate of score reproducibility for many achievement tests.

Pass-fail decision reproducibility is the more important reproducibility, especially for examinations that have serious consequences associated with passing or failing – examinations such as end-of-curriculum tests, licensure and certification examinations or pre-employment tests which must be passed in order to gain employment. Several approaches are used to estimate the reproducibility of pass-fail decisions, but one convenient procedure was suggested by Subkoviak (1988), which uses tabled values associated with the score reliability coefficient and a *z*-score calculated from the passing score on the test and the mean and standard deviation of the test.

Correspondence of test scores with achievement

Cognitive achievement is measured on a continuum of knowledge and skills acquired by the examinee. A fundamental assumption of all testing is that test scores correspond (within measurement error) in a one-to-one fashion with the knowledge attained by the examinee. High scores indicate high mastery of the subject and low scores symbolize low attainment of the content tested. Errors of measurement (low reproducibility) reduce the confidence one has in this one-to-one correspondence, since measurement error introduces noise into the system and clouds one's ability to make appropriate and correct inferences. Low score reproducibility or reliability thus reduces the validity evidence for an examination, since inferences to the domain of interest are made more questionable.

Thus, score and pass-fail reproducibility are an essential source of validity evidence for most tests. "Reliability is a necessary but not a sufficient condition for validity" is an often cited phrase in the educational measurement literature. Test scores must be reliable or reproducible in order to have any chance of being valid but just because tests are reliable does not necessarily mean that they are valid. Another familiar phrase, especially in medical education, is that some testing methods trade off reliability for validity. This is, of course, nonsense if one accepts that reliability is an essential source of a test's validity evidence.

Defensibility

All test scores must be defensible. At a minimum, a teacher must be able to defend the keyed correct answer to students. At the other extreme, a high-stakes licensure or certification agency must be able to defend its test scores and its pass-fail decisions in court.

There are many important aspects to examination defensibility. Some characteristics will be more important than others for certain types of inferences. But the one common thread running through all defensibility arguments for examinations is *validity evidence*. Test validity evidence, as discussed above, refers to all data and information assembled to support or refute the argument for the legitimacy and reasonableness of the inferences made from the examination scores.

http://www.springer.com/978-1-4020-0466-7

International Handbook of Research in Medical Education

Norman, G.R.; van der Vleuten, C.P.M.; Newble, D.I. (Eds.)

2002, XIII, 1106 p. In 2 volumes, not available separately., Hardcover
ISBN: 978-1-4020-0466-7