2011; 33: 447–458 MEDICAL TEACHER

Post-examination analysis of objective tests[1]

MOHSEN TAVAKOL & REG DENNICK
University of Nottingham, UK

Abstract

One of the key goals of assessment in medical education is the minimisation of all errors influencing a test in order to produce an observed score which approaches a learner's 'true' score, as reliably and validly as possible. In order to achieve this, assessors need to be aware of the potential biases that can influence all components of the assessment cycle from question creation to the interpretation of exam scores. This Guide describes and explains the processes whereby objective examination results can be analysed to improve the validity and reliability of assessments in medical education. We cover the interpretation of measures of central tendency, measures of variability and standard scores. We describe how to calculate the item-difficulty index and item-discrimination index in examination tests using different statistical procedures. This is followed by an overview of reliability estimates. The post-examination analytical methods described in this guide enable medical educators to construct reliable and valid achievement tests. They also enable medical educators to develop question banks using the collection of appropriate questions from existing examination tests in order to use computerised adaptive testing.

Introduction

The purpose of this Guide is to provide an overview of the rationale and processes involved in analysing and evaluating the results of objective assessments in medical education. By objective assessment we mean multiple choice questions that assess knowledge and objective structured clinical examinations (OSCEs) and related assessments (e.g., direct observation of procedural skills (DOPS), mini-clinical examination (mini-CEX)) that assess clinical skills by means of specific and easily measurable observational criteria. We are consequently excluding material such as essays, assignments or portfolio-based assessments which rely on more subjective interpretations of performance. We acknowledge that we are concentrating on analysing the results of measuring things which are 'easier' to measure and that therefore we are guilty of bias. There are many learning outcomes of medical education in the affective or attitudinal domains that are notoriously difficult to measure objectively but which are exceedingly important. Nevertheless the objective testing of knowledge and clinical skills is a major element of medical assessment and an understanding of the processes whereby these measurements are made, analysed and evaluated is an essential requirement of contemporary practice. A number of text-books and papers have covered this important area (Traub & Rowley 1991; Gilbert 1996; Anastasi & Urbin 1997; Hopkins 1998; Osterlind 1998; McAlpine 2002; Shultz & Whitney 2005; Crocker & Algina 2008; Holmbow & Hawkins 2008; Rust & Golombok 2009; de Champlain 2010; Cohen & Swerdlik 2010).

From the outset we assert that objective testing is a form of measurement, termed *psychometrics*, conceptually related to the principles of measurement in general. Consequently factors such as accuracy, reliability, reproducibility, validity, specificity and sensitivity can all apply in varying ways to the process of objective measurement. The control of these factors is made more important by the fact that psychometrics applies to human beings with all their intrinsic propensity for variation. A physical property such as length or mass can be measured extremely accurately whereas the measurement of human learning is associated with significant variation and 'noise'. In addition, in the case of learning, it is clear that it is not a homogeneous entity. Traditionally (Bloom 1956) said that it is differentiated into the cognitive, psychomotor and affective domains with further hierarchical levels within each. In this guide we will concentrate on the knowledge domain as measured by multiple choice questions and some aspects of the psychomotor domain measured by OSCEs.

Practice points

- Medical educators need to measure how much material has been learned by evaluating the results of particular achievement tests.
- Analysing examination questions and scores by means of psychometric methods improves objective tests.
- Both quantitative and qualitative approaches examine how individual test questions function in an assessment.
- Item analysis provides evidence for exam questions that need to be adapted, revised or discarded.
- Reliability analysis reveals the consistency, usefulness and practical value of a test.

Correspondence: Professor Reg Dennick, Medical Education Unit, The University of Nottingham, Nottingham, NG7 2UH, UK. Tel: 0044 (115) 823 0013; fax: 0044 (115) 8230014; email: reg.dennick@nottingham.ac.uk

ISSN 0142–159X print/ISSN 1466–187X online/11/060447–12 © 2011 Informa UK Ltd.
DOI: 10.3109/0142159X.2011.564682

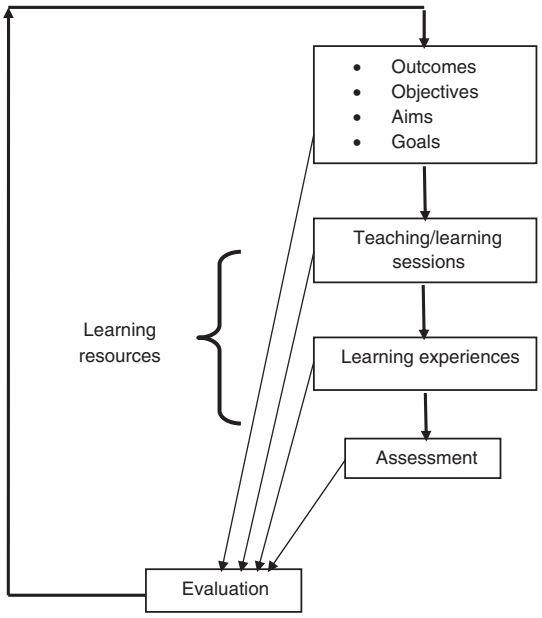

Figure 1. Constructive alignment.

A further point of clarification is that although the terms measurement and evaluation are often mistakenly tossed into one terminological basket, each has a distinct meaning and should be differentiated from each other. Measurement is the process of assigning a numerical value in order to assess the magnitude of the phenomenon being measured. For Ralph Tyler, evaluation refers to 'the process of determining to what extent the educational objectives are being realised' (Tyler 1949). Obtaining and reporting a numerical index has no meaning in itself unless we interpret and value the index (Morrow et al. 2006).

The measurement of learning is not an isolated event; it is fundamentally part of a curriculum cycle beginning with the identification of the learning needs and then the learning outcomes of, for example, doctors, nurses and other healthcare professionals. These learning outcomes then become the basis of decisions made concerning the most appropriate methods of acquisition, such as lecturing, small group teaching or experiential learning. It is after this point that the measurement of learning or *assessment* takes place to see if the required outcomes have been achieved. It is here that the subject of this guide, post-examination analysis, takes place but this is just one component of an overall process of curriculum *evaluation* that seeks to ascertain if the curriculum elements of outcomes, learning and assessment are articulated optimally in what has been termed *curriculum alignment* (Biggs & Tang 2007). This concept is summarised by the diagram in Figure 1.

Curriculum designers and teachers should have a clear and collective picture of learning outcomes. They are statements of what students are expected to learn and demonstrate by the conclusion of the learning process and in principle they need to be measurable and hence capable of being transformed and operationalised into objective assessments.

Although this Guide will focus on the methods for analysing the data generated by assessments it must be realised that the information obtained from this data feeds back into the processes of learning, teaching and outcome specification. For example, anomalies revealed in tests might indicate poor question setting, poor teaching, or even the specification of inappropriate learning outcomes. The assessment cycle can be displayed diagrammatically as shown in Figure 2.

The examination cycle

For the purposes of this Guide we will assume that learning outcomes have been defined and that appropriate teaching and learning experiences have been provided so that these outcomes can be achieved by all learners. There is some controversy still surrounding the terminology associated with learning outcomes. Outcomes are meant to be broad statements describing the competencies required or achieved by learners at the end of a course of study. For example 'outcome' based medical curricula have been defined by the GMC, Scottish Deans and the WFME (GMC 2003; WFME 2003; Scottish Dean 2007).

'Learning objectives', on the other hand, are more granular and are frequently used to describe the learning that has been acquired at the end of a specific learning episode such as a lecture. Whatever level of granularity is specified outcomes or objectives are statements describing what learners should be able to do. For simplicity we will use the term outcomes throughout.

As previously pointed out, learning outcomes should be measurable and hence they are frequently termed behavioural outcomes. Bloom (1956) classified behavioural outcomes into three domains: the cognitive domain, the affective domain and the psychomotor domain. Within the cognitive or knowledge, domain outcomes can be categorised on a spectrum of increasing cognitive demand. Bloom's original ranking was differentiated into the following: knowledge, comprehension, application, analysis, synthesis and evaluation. More recently the knowledge dimension has been updated, giving the following: remembering, understanding, application, analysis, evaluation, creation (Anderson & Krathwohl 2000). The original psychomotor domain (Simpson 1966) consisted of general terms which were not easy to operationalise into an observational protocol: perception, set, guided response, mechanism, complex overt response, adaptation and origination. The Dreyfus model (Dreyfus & Dreyfus 2000) is now widely used to monitor the acquisition of practical skills but again is not easily transformed into an objective system for assessing, for example, practical procedures. As will be seen methods for assessing at OSCE stations essentially revolve around defining a list of specific practical competencies that can be easily observed rather than measuring an individual against a scale of increasing psychomotor complexity. Measuring outcomes in the affective domain is achieved by observing defined behaviours but the criteria are often subjective and difficult to define. The relationship between observed behaviour and an individual's internal 'attitude' is also problematic.

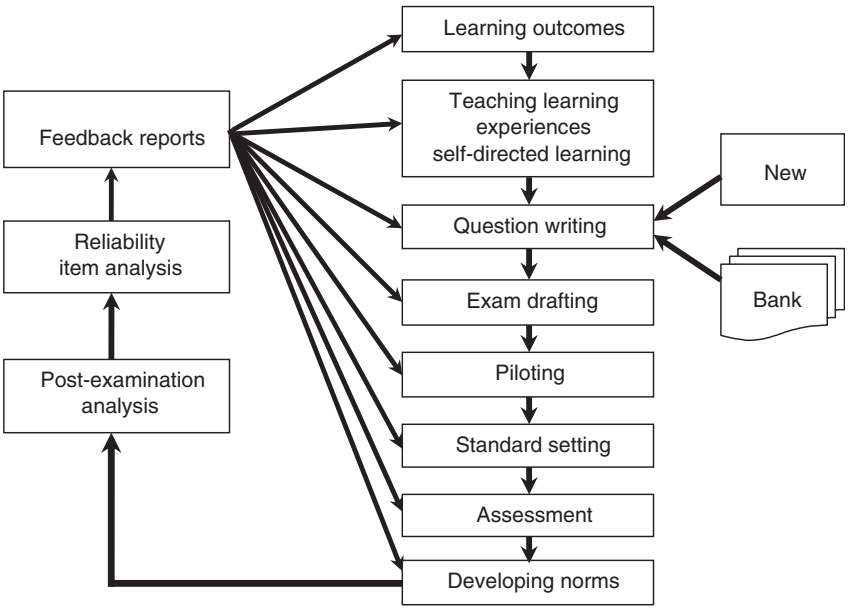

Figure 2. The examination cycle.

Test

Test has been defined as 'an objective and standardised measure of a sample of behaviour' (Anastasi & Urbin 1997). To understand this definition, three key elements need to be clarified, that is objectivity, standardisation and a sample of behaviour. A test is considered to be objective if it is administered, scored and interpreted independently of the subjective judgment of examiners. A standardised test is a test in which the procedure for the questions, scoring, interpreting and administrating are 'uniform from one examiner and setting to another' (Gregory 2007). This simply shows if we want to compare students' scores to each other, it is necessary to test all students with the same test questions under the same test conditions. A test should measure a sample of particular behaviour. Based on this sample, test constructors draw some inferences and hypotheses. For example, if a medical educator wishes to test the knowledge of students' medical terminology, he or she examines their performance with a representative sample of medical terms.

Item writing and item banking

If learning outcomes have been defined and appropriate learning experiences provided so that students can acquire them, the next phase becomes the writing of test items, questions or OSCE check lists. Question writing is a skill that needs to be carefully developed as badly constructed items can subvert and damage the alignment of the assessment process. Questions should be related to a defined learning outcome whose cognitive level is known and they should be clear and unambiguous. Questions should also be valid, i.e. they should have at least content validity, construct validity and face validity, concepts defined in Hopkins (1998). There is not space here to go into the techniques of question writing but the reader is referred to Case and Swanson (Case & Swanson 2010). It would be useful if question developers produced twice as many questions as required in the final draft. Question developers can keep these questions in the question bank for later decisions on incorporation into the test. The test developer could also design parallel forms of the test in order to determine reliability by using parallel-forms reliability estimation (see reliability estimates). In addition to question writing, assessors are increasingly turning to question banks where many questions have been constructed, tested and evaluated and refined, for example the Universities Medical Assessment Partnership (UMAP), at Leeds Institute for Medical Education (UMAP 2010). The Hong Kong Ideal Consortium has also created and shared an assessment bank for medical educators on an international scale (Ideal Consortium 2010). The advantage of a large bank of questions means that assessors have convenient access to a large number of tried and tested questions which are categorised according to the target content area, psychometric properties or other independent variables. Question banks can be stored in computers and used for computerised adaptive testing (CAT) where the question delivered to the student is a function of their performance on the previous questions (Weiss & Vale 1987). In this approach the programme may not allow the student to move on to the next question, if he or she does not correctly answer the previous question. This is very useful for formative assessment or in high stakes examinations, where institutions are deciding if a candidate will be certified or licensed (Bergstrom & Lunz 2008). CAT operates by measuring the performance level of the student during the test. After each question, his/her current performance can be compared to all questions in the bank. The algorithm of the computerised testing programme selects the next question from the bank based on the current level of the student's performance and all test specifications. This process continues until the test is terminated. By this method the questions that are too easy or too difficult will not be delivered to that candidate and the test will be individualised. Using CAT, the numbers of test questions that need to be administrated are reduced by 50%

without sacrificing reliability and concurrently the measurement of error is reduced by 50% (Bergstrom & Lunz 2008; Cohen & Swerdlik 2010).

Item sampling: how many questions should we ask?

When we set an exam it is practically unfeasible to ask a question concerned with every single learning outcome in an area of learning, consequently we are forced to sample for practical reasons. However, if a particular area of knowledge has been described by a range of learning outcomes, which cover an appropriate depth and breadth of the domain, what fraction of these learning outcomes constitutes a representative sample from the total population? In other words how many items should we set in the test to reassure us that the score obtained for a student reflects their global knowledge? This is a question that is not often asked as the size of many exams is based on tradition or length of time rather than appropriate sample size. The validity created by addressing this issue is associated with content validity.

The process of selecting a representative fraction of a total pool of items is referred to as item sampling. The size of items in a test can be a source of error and error leads to unreliability as will be discussed later (Cortina 1993). However, it is clear that as the number of test items increases sampling error will decrease and hence reliability should increase as shown in Figure 3. In addition, in multiple choice tests where there is the possibility of guessing, increasing the number of items will reduce errors associated with guessing.

An appropriate sample size can be calculated for a test using the formula below:

$$n = \frac{Z^2(SD)^2}{e^2}$$

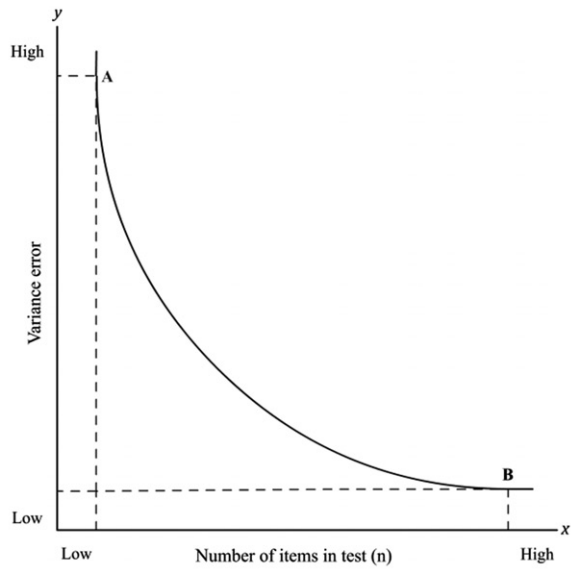

Figure 3. Relationship between the number of items in the test and variance error.

Where n is equal to the sample size, Z^2 is a confidence level indicating how much the sample size is influenced by chance (1.64 for 90% confidence, 1.96 for 95% and 2.57 for 99%), *SD* is an estimation of standard deviation in the population of items, e^2 is the error of the sample size, e.g., 0.03 or 0.05. To calculate the sample size we therefore need the standard deviation estimated either from a pilot study or from previous data. For example if a random sample of 20 questions is drawn from the population of items and piloted with a group of students a standard deviation of student's scores might be calculated as 0.26. This standard deviation is then substituted into the sample size formula and the sample size required is calculated. With this information in hand, if we desire to obtain a sample with a 95% confidence level and 0.05 precision, the sample size required (n) is calculated as:

$$n = \frac{(1.96)^2(0.26)^2}{(0.05)^2} = 106$$

Therefore, the test should include at least 106 questions to provide 95% confidence. It should be noted that if the heterogeneity of the population of items being tested is large, a larger sample is required to obtain a given level of precision and vice versa. The more heterogeneous a test, the less inter-item consistency there is as determined by the Cronbach alpha statistic as discussed later.

Piloting of assessments

In principle, once an examination has been conceptualised and constructed it should be piloted on an appropriate group of learners to eliminate any gross content problems using item analysis. Indeed, statistical procedures assist us to judge 'good' questions and those questions that need to be modified or discarded. On the basis of this analysis, test questions are created and tried out on a new sample of examinees to finalise the test. However, in practice this process may be unfeasible as it might allow 'leakage' of the exam content into the student body and might use up valuable questions from the bank. An alternative or additional approach is to ensure that the exam paper is seen by an external examiner who can give valuable advice on the ease, difficulty or appropriateness of questions. Another consideration is the use of parallel-forms of the test, which is discussed below.

Standard setting

Once the exam paper is ready for use a pass-mark needs to be determined by 'standard setting'. Briefly a group of appropriate individuals examines each question in turn for its ease or difficulty in relation to its target audience and, via a subjective process of consensus, establishes trustworthy, justifiable and acceptable standard scores for both written tests and clinical examinations. The standard is the minimum adequate level of performance, indicating the boundary between those who perform acceptably and those who do not (Norcini 2003). There are a number of different standard setting methods available for identifying the standard score, described in the literature (Cusimano 1996, Norcini & Guille 2002; Bandaranayake 2008), but it is not the purpose of this Guide

to discuss them. However, what is relevant here is the means by which the results of post-examination analysis can be fed back to those involved in standard setting so that their ability to evaluate the attributes and quality of questions is enhanced. Standard setters need to build up a body of knowledge and experience of items and overall test construction so that they can make more informed judgements concerning the ease or difficulty of a test and hence define a fairer and more appropriate pass mark or cut-score.

Norm referenced and criterion referenced tests

One final topic that needs to be considered is the overall purpose of the test since this will affect the number and type of questions it contains and hence will influence the interpretation of post-examination analysis. The two major purposes for constructing tests are for norm-referencing and criterion referencing, terms coined by Robert Glaser (Glaser 1963).

In norm-referenced approaches, test designers seek to differentiate high-performing students from low-performing ones in order to select the best students for particular reasons, such as a limited number of places on future courses, for competitive reasons or to maintain particular standards. At the conclusion of the examination test makers calculate the mean and standard deviation of students' test scores and then determine the placement of each student on a normal distribution curve. Arbitrary grade boundaries and pass-marks are inserted onto this distribution depending on the proportion of students that are permitted to pass and the availability of places for students with different grades. Clearly how well a student does in a norm-referenced exam is a function of how the whole cohort performs rather than being an absolute measure of the student's attainment. If the purpose of the exam is competition for limited places or prizes then a norm-referenced exam should be designed to spread students out along the normal distribution. Thus, in a knowledge based exam it would be appropriate to have a range of questions with heterogeneous cognitive demand and post-examination analysis would be able to confirm whether such an exam achieved its aim.

Criterion-referenced approaches attempt to assess learners by measuring their attainment without reference to the performance of others. Criterion-referenced assessment, according to Cohen and Swerdlik (2010), is defined as 'a method of evaluation and a way of deriving meaning from test scores by evaluating an individual's score with reference to a set standard'. Thus, whether a student passes or fails or achieves a particular grade is determined by their ability to achieve a particular score in an examination regardless of the attainment of the cohort. Criterion-referencing is used when the primary purpose of an examination is to see if students have attained specific cognitive or psychomotor competencies. Clearly this will influence the number and type of questions asked and it is more likely, therefore, that a criterion-referenced exam would be more homogeneous. Nevertheless given the range of abilities within a population of learners and the residual heterogeneity of questions, even in a criterion-referenced exam, it is still likely that a normal distribution will be observed when the results are analysed, albeit with less variance.

Analysis of examination questions

The rationale for using post-exam analysis techniques is to improve the quality and reliability of assessments, and to select the questions that are most appropriate for assessing students in order to estimate students' level of performance with low variance error. This section will explain how various processes can be used to identify flawed questions, how discrimination can be improved and how overall reliability can be increased by deleting or rephrasing questions. In the case of OSCE examinations the identification of 'hawks' and 'doves' and the problem of dealing with inter-rater reliability issues will be discussed.

In addition we will discuss how post-exam analysis can feed into the development of meta-data coded questions in item banks and how question data can influence the examination cycle by, for example, improving the accuracy of standard setting.

Moderation

Before dealing with the more analytical methods it is worthwhile briefly discussing the process of exam Moderation. This is more likely to be required in situations when there are multiple assessors who are not necessarily using very objective assessment criteria. This is unlikely to occur with machine or computer marked knowledge based assessments but can occur with human assessed OSCE examinations. Examiners, especially when subjectivity is involved in test scoring, who differentially interprets assessment criteria ('hawks' and 'doves') can be source of error variance, which in turn may negatively or positively influence the student's true score on the performance being measured. This will also generate unreliability and mechanisms need to be in place to compensate for this factor (see later). For written examinations using essays or short answers, double, anonymous marking is often the best way to ensure that fair standards are maintained.

Descriptive statistics

Once raw scores have been obtained from a test the simplest analysis that can be undertaken is to look at the frequency distribution of scores and to calculate the mean, the mode, the median and the standard deviation. These figures are readily calculated by inserting the data into SPSS (SPSS 2009). The definitions of the commonly used terms are described in Table 1.

Inspection of the distribution can reveal how far it deviates from a 'normal' distribution and how skewed it is. Differences between the mean, mode and median also give a more objective indication of how much the distribution deviates from normal. How skewed the distribution is can indicate the overall ease or difficulty of a test. When the mode is off to one side the distribution is said to be skewed. If the mode is to the left with a long tail to the right the distribution has positive or right skewness. This shows that few students' test scores fall at

Table 1. Definitions of some commonly used terminology.

- The mean is the average of all the scores.
- The median is the midpoint of the distribution, where 50% of the scores fall on either side.
- The mode is the most frequently occurring score in the distribution.
- The deviation is the distance between an individual's score and the mean score.
- The variance is the average, squared, deviance from the mean.
- The standard deviation is the square root of the variance and gives an indication of the spread and variability of the data.

the high end of the distribution, which means the test was too difficult. Here, in order to modify the discrimination at the lower end of the distribution, more questions with a lower level of difficulty should have been used. If the long tail is to the left the distribution has negative or left skeweness. This shows that few students' test scores fall at the lower end of the distribution, which means the test was too easy. Here, in order to modify the discrimination at the higher end of the distribution, harder questions could have been used.

Z-scores

The means and standard deviations of raw scores will vary depending on the size of the examination and the total mark. A method for 'normalising' examination scores so that they can be compared in a standard way is to calculate *z*-scores.

If students' scores have a mean and standard deviation, then the student's score is transformed into a *z*-score by the equation:

$$z = \frac{X - \bar{X}}{S}$$

This formula simply shows that a *z*-score is equal to the difference between a raw score (X) and the mean score of students $(\bar{X})$ in a particular test divided by the standard deviation (s). All *z*-score transformed distributions have a mean of 0 and a standard deviation of 1. An individual's *z*-score shows how far above or below the mean their score is in units of standard deviation. For example assuming the mean of scores in a particular test is 50 with a standard deviation 15, if a student scores 65, his/her z score is +1. This means that the student is +1 standard deviation above the mean of the distribution. Standard tables of *z*-scores are available for comparing the position of student to each other. Within the normal distribution the position of scores is as follows: 68% of scores lie within +/−1 standard deviation of the mean. 95% of scores lie within + /−2 standard deviations of the mean. Finally, 99.75% of the scores lie within +/−3 standard deviations of the mean. Therefore, in the above example, approximately 16% of other students obtained higher scores than the student. Thus, relying on a raw score can provide a wrong impression of the student, as well as a distorted view of the exam. *Z*-scores allow teachers to compare students' scores on different tests with different total marks.

Item analysis

Item analysis of test results uses quantitative methods to help make judgements about which questions need to be adopted, which questions need to be revised and which questions should be discarded. Item analysis can be used to analyse the ease or difficulty of individual questions as well as the relationship between individual questions and the global test score. For example if a student with a high score on the test answers a question correctly that question would be considered a good question. Equally, if a student with a low score on the test answers a question incorrectly, that question would be considered a good question. On the basis of item analysis, a revision and improvement of the test can be made (Cohen & Swerdlik 2010).

The item-difficulty index

If all students answer a particular question either correctly or incorrectly, that question is not a good question and needs examining. It is either too easy or too difficult. Questions which are too easy or too hard for a student contribute little information regarding the student's ability (Green et al. 1984). The item-difficulty index (sometimes also termed item facility) refers to the percentage of the total number of students who answered the test question correctly and is calculated as follows:

$$P_i = R_i/N$$

Where R is the total number of students who answered the question correctly and N is the total number of responses (correct + incorrect + blank response). The question number is called i. P is the fraction of correct answers. For instance, if 40 of the 100 students answered question 1 correctly, the item-difficulty index is simply calculated as follows:

$$P1 = 40/100 = 0.40$$

Consequently the value of an item-difficulty index ranges from 0 (if no one answered a question correctly) to 1 (if everyone answered a question correctly). In addition, the larger the P value, the easier the question. If P_i is located between 0.3 and 0.8, the question is considered to be a good question.

However, the effects of guessing in objective tests needs to be considered. For example, the probability of guessing a five-option multiple choice questions is equal to 0.20. In order to calculate the corrected question difficulty in this case, we need to add the probability of getting an answer right and 1.00 (if everyone answered a question correctly) and then divide the sum by 2, that is, $(0.20 + 1.00)/2 = 0.60$. Conversely, the corrected question difficulty in a true-false question, where the probability of a correct answer is 0.5, is equal to 0.75, i.e. $(0.5 + 1.0)/2$. In these cases 'good' questions should have a range of item difficulty between 0.2–0.6 and 0.5–0.75, respectively.

Davis has presented the following general formula for determining item difficulty when questions need to be corrected for 'chance success', (Davis 1952):

$$P_{Di} = 100 \times \frac{R - \frac{W}{n-1}}{K - KR}$$

Where P_{Di} equals the item-difficulty index corrected for the question number (i), R equals the number of students who answer a question correctly, W equals the number of students who answer a question incorrectly, n equals the numbers of choices in a question, K equals the number of students, and KR equals the number of students who do not answer the question.

The item-discrimination index

The item-discrimination index is a value of how well a question is able to differentiate between students who are high performing and those who are not, or between 'strong' and 'weak' students. The item-discrimination index is symbolised by a lowercase italic "d". The range of d is −1.00 to 1.00. The most common method to calculate a d-value for individual questions is now described.

In this method, the examiner divides students into two groups ('high' and 'low') according to the score sheet of each student. On the basis of this classification, 27% of the students are categorised as a strong group and 27% as a weak group. Some methods prefer a *'top third'* and a '*bottom third*' but studies have shown that when students are divided into two groups on the basis of 27% the sensitivity and precision of the value of d is increased (Kelley 1939; Cohen & Swerdlik 2010). Obviously 46% percent of the middle-scoring students are excluded from the calculation of the item-discrimination index.

Next the number of students (in both groups) who answer a particular question correctly is calculated. The following formula is used to calculate a d-value.

$$d = \frac{(U - L)}{n}$$

Where U equals the number of correct answers in the upper group, L equals the number of correct answers by the lower group and n is 27% of the total number of students (Cohen & Swerdlik 2010). For example, a physiology lecturer gave the neurophysiology test to a total of 112 medical students. The lecturer isolated the top and bottom 27% of the test scores, with a total of 28 students in each group. The lecturer observes that 18 students in the 'strong' (top) group answered question 1 correctly and 10 students in the 'weak' (bottom) group answered correctly question 1. Therefore, the d-value is equal to 0.28 = [(18–10)/28]. The higher the d value, the better and more discriminating the test question.

If a given question has a high d-value, it is likely to be very discriminating. However, a negative d-value on a given question indicates that the 'strong' students answered the question incorrectly and the 'weak' students answered the question correctly. Such questions should either be revised or discarded.

The point bi-serial correlation coefficient

Another widely used method for computing the item-discrimination index is the point bi-serial correlation coefficient. This is a statistic that indicates the relationship between a particular question (correct or incorrect) on a test and the total tests score (Kaplan 2008). Questions are scored 1 for 'correct' and 0 for 'wrong'. The sum of correct answers of test questions produces the total student's score. To calculate the item-discrimination index for each question the following formula is used:

$$R_{pbi} = \frac{\bar{x}_r - \bar{x}_w}{S_t}\sqrt{P(1-P)}$$

Where R_{pbi} equals the point bi-serial correlation coefficient, $\bar{x}_r$ equals the mean score of students answering the item correctly (those coded as 1s), $\bar{x}_w$ equals the mean score of students answering the item incorrectly (those coded as 0s), and S_t equals the standard deviation for the entire test. P equals the proportion of students answering the item correctly. The higher the R_{pbi} value, the better the question is at discriminating. The R_{pbi} values range from −1.0 to +1.0. A question with a negative R_{pbi} should be revised or discarded.

Statistical significance

The last method for the assessment of the quality of a multiple-choice question is to divide students into two groups, those who answered the item correctly, called 'group R', and those who answered the item incorrectly, called 'group W'. The mean of the total score of 'group R' and 'group W' is calculated. The mean score of group R ($\bar{X}_R$) could be below or above the mean score of group W ($\bar{X}_W$). Consequently, the null hypothesis that should be considered is that '$\bar{x}_R$ is equal to $\bar{x}_W$', weighed against the alternative hypothesis that '$\bar{x}_R$ is greater than $\bar{x}_W$'. The null hypothesis means that there is no difference between the mean scores of the students who chose the correct answer and the wrong answer. To test the null hypothesis a t-test can be used that assesses whether the means of two groups ($\bar{X}_R$ and $\bar{X}_W$) are statistically different from each other. If the p-value is less than 0.05, we will reject the null hypothesis and accept the alternative hypothesis. This means that the test question has divided students into two separate strong/weak groups.

Reliability

The main methods of calculating reliability will be described and explained. Examples of the use of point-biserial correlation co-efficient and Cronbach's Alpha will be used to explain how the reliability of tests can be improved. The evaluation of inter and intra-rater reliability in OSCE examinations will be described.

The traditional way of explaining and defining reliability is that it is concerned with the reproducibility, stability and internal consistency of an assessment. In the psychometric literature, reliability more generally refers to the consistency of a measurement tool (Cohen & Swerdlik 2010). For example a test is stable if a student repeatedly takes the same test and obtains the same mark. Reliability is a measure of a test's ability to generate similar results when applied on different occasions. When the difference of scores obtained by the same test on different occasions is high, the test cannot be reliable and is fatally flawed. For example, if the same achievement test delivers scores for a student of 62, 85, 53 and 92 in different

points in time, then this test is not consistent and needs to be investigated. The internal consistency of a test is a measure of how well the individual items are functioning together to measure the same underlying constructs and how accurately and precisely can it measure the construct of interest.

Another way of thinking about test reliability is that it is a function of the difference between the observed test score of the student and his/her 'true' score. The observed score is the score that a student obtains from an actual test. The true score is the score that a student obtains from a (hypothetical) test when it accurately measures his or her underlying ability. If there is a significant difference between an observed test score and a true score, the reliability of the test is low, and vice versa.

However, a more general view of reliability is that it is concerned with the *error* inherent in psychometric measurements. In the language of assessment, there are two main factors that cause errors in measurements, external and internal factors (Anastasi & Urbin 1997). The external factors depend on the test situations and administrations, such as the room temperature, guessing answers, emotional problems, physical discomfort and lack of sleep. The internal factors depend on the quality and quantity of the test, such as item sampling (the limited number of test items) and the way in which the item is constructed. Scorers and scoring systems can also be a potential source of error.

Classical test theory proposes that an individual possesses a particular amount of, for example, knowledge, given the symbol T for 'true score'. However the measurement of this score, X, or the 'observed score' is confounded by the errors of measurement, E. Thus,

$$X = T + E$$

The task facing the designer of high quality assessments should be to identify and minimise these sources of error. Physical measurements of size or temperature may have relatively limited and easily controlled errors of measurement whereas cognitive, psychomotor or affective measurements on human beings may have large, uncontrollable and even unknown errors. The three sources of error influencing reliability derive from: the test, the testee (student) and the tester. In many cases errors can be identified and controlled before an assessment is undertaken but it is practically impossible to estimate every possible error. As a result, the determination of the true reliability coefficient of a test is not practicable. Therefore, medical teachers have to estimate the reliability of a test from the data acquired after the test has been administered using the techniques for estimating reliability to be described below.

The test

The Test can be a written knowledge-based MCQ, an OSCE station or some other form of psychometric assessment. Errors are created in its production and interpretation and by processes impacting on the testing environment:

- Ambiguous questions
- Too long (fatigue)/Too short
- Invalid questions
- Non-homogeneous question paper
- Too hard/too easy
- Poor instructions
- Too hot/too cold/too noisy
- Not enough time
- The level of lighting
- Responses which are coded incorrectly

The tester

The **tester** can be the person responsible for creating a written test such as an MCQ or, in the case of an OSCE, or other practical assessment, the person who is responsible for using and interpreting the assessment criteria. Errors can be created from a lack of understanding of assessment principles or item construction or by a lack of training in applying assessment criteria:

- Lack of understanding of learning objectives
- Poor interpretation of assessment criteria
- Inconsistent application of assessment criteria
- Inconsistent scoring system or mark scheme
- Sexist/racist bias
- Systematic typing errors
- Lack of assessment training
- Inter-rater variability
- Subjectivity in scoring

The testee

The testee is the person being tested. Error and variation not due to the intrinsic capacity of the individual can be caused by their reaction to stress or illness or by a lack of appropriate teaching or learning preparation:

- Stress
- Therapy and illness
- Lack of teaching
- Inconsistent teaching
- Poor learning environment
- Lack of appropriate resources
- Lack of practice opportunities
- Lack of sleep

Reliabilities estimates

Test–retest reliability estimates

The test-retest reliability is estimated by performing the same test at different times with the same students. The correlation coefficient (r_a) between students' scores in the two tests is used as a quantitative measure of the test–retest reliability.

A limitation of the test–retest reliability is that the passage of time can influence the student's response in the second test. This is because students may learn new things, forget some things or acquire new skills.

Parallel-forms reliability estimates

The use of parallel forms of the test helps avoid the difficulties involved in test–retest reliability. To estimate the reliability, two different samples of a test are tested on two different occasions. In the first occasion, students are tested with one form of the test. On the second occasion, the same students are tested with a parallel form of the test. Parallel test forms should have the same average item difficulty. In addition, the mean and the standard deviations of observed test scores in parallel forms should be equal. Estimating parallel forms reliability is similar to estimating test–retest reliability. Students' scores on the two test parallel forms are correlated to obtain an estimate of parallel-forms reliability.

One obvious limitation of estimates of parallel-forms reliability is that test scores may be affected by factors such as fatigue, motivation and learning.

Split-half reliability estimates

To estimate the reliability of a whole test, a single test can be administrated followed by a splitting of the items into halves; odd-numbered items to one half of the test and even-numbered to the other half. A Spearman–Brown correlation can also be used to estimate the effect of shortening the test on the reliability coefficient. Reducing the size of a test appropriately can minimise administration time and students' exam fatigue (Cohen & Swerdlik 2010).

Coefficient alpha

The value of alpha is widely used for estimating the internal consistency reliability or item homogeneity of a test (Henson 2001; Streiner 2003). In contrast to previously described reliability indices the test is only administered once and the scores used to calculate the alpha statistic. Alpha can be considered as an estimate of the interrelatedness of a set of test's items (Schmitt 1996). The value of alpha also indicates how similar or unique test items are (Cortina 1993). The alpha coefficient can be used on either tests with dichotomous or polytomous items. It ranges in value from 0 to 1 and should be above 0.70, but not much than 0.90 (Nunnally & Bernstein 1993; Streiner 2003). As the obtained score is a reflection of all the test's items, examiners seek for a high value of alpha on a test. However, although a high value for alpha is usually better, it is not always the case. Tests that measure a single domain can contain some degree of heterogeneity among the items. If a test taps a single domain but has three or four sub-domains, the homogeneity of each sub-domain can more than the test as a whole. If this is the case, the value of alpha should not be over 0.90. In this situation a large value of alpha is an indication of 'unnecessary duplication of content across items and points more to redundancy than to homogeneity' (Streiner 2003).

Kuder–Richardson reliability. As previously pointed out higher test homogeneity generates a higher internal consistency within a test. The most common statistical procedure for estimating the internal consistency reliability in achievement tests is Kuder–Richardson 20 (KR-20). In contrast to coefficient alpha, KR-20 is used to determine the internal consistency of dichotomous items such as objective tests which can be scored as either right or wrong. The value of KR-20 is directly proportional to the strength of the relationship between the items on the test. It ranges between 0 to 1 where 0 represents a lack of reliability and 1 represents a fully reliable test.

In summary to estimate the impact of item homogeneity on test reliability, two different indices of internal consistency reliability are available: coefficient alpha and KR-20. It should be noted that a lower reliability value indicates that all the items on the test measure a diversity of knowledge or performance. Furthermore, the reliability index is affected by the test and students' heterogeneity. Longer tests and heterogeneous students will have a higher internal consistency reliability (Anastasi & Urbin 1997).

Psychometric properties of OSCE

The psychometric analysis of OSCE stations has been less reported in the medical education literature in comparison to knowledge based tests. Depending on the purpose of the examination, the number of stations can vary and each station can assess a specific ability of the candidate. To quantify a specific behaviour, checklist items, which correspond to specific actions, are objectively devised by content experts through consensus. The examiner marks the student in each station by checking whether or not a given action was performed competently either dichotomously or on a scale. At the end of each station, examiners record their scores and feedback on the performance of students. The number of items in each station can vary. As an example, in station 1 with 21 items on the checklist, a student might competently perform 15 clinical actions. Therefore, he or she receives a total score of 15 out of 21 from station 1. If the OSCE consists of 25 stations and uses a rating scale for measuring student performance 25 ratings are calculated and then the mean for each student. Before the OSCE, the overall pass mark for each station is decided by standard setting. Other assessment procedures might include a global judgement by the examiners of pass, fail or borderline given independently of any scoring.

Station analysis of OSCEs

In the OSCE, each station is regarded as an item of analysis. The first common analysis is to determine the inter-station reliability of the OSCE which refers to the degree of correlation between all the stations on the OSCE. Calculating the index of inter-station reliability is useful in assessing homogeneity. OSCEs are homogeneous if they contain stations that measure a single trait.

The Kuder-Richardson 20 formula (KR-20) allows medical teachers to estimate inter-station reliability. The KR-20 formula is

$$r_{KR20} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma pq}{\sigma^2}\right)$$

Where r_{KR20} provides a reliability coefficient of the whole OSCE, K is the number of stations in the OSCE, σ^2 is the variance of total station scores, p is the proportion of students

who pass the station, q is the proportion of students who fail the station ($q = 1 - p$), and Zpq is the sum of the pq products over all stations. K-R20 is calculated using SPSS. The higher the reliability coefficient, the more homogenous the OSCE.

A low reliability coefficient shows that a number of stations are performing poorly in assessing the clinical competencies contributing to the OSCE examination.

If the reliability coefficient is low it suggests that some stations do not share equally in the common core clinical performance and need to be revised or discarded. Therefore it is important to detect them by computing the correlation of each station with the total OSCE score. This involves using the point-biserial correlation method as previously described. The item (station) total correlation test allows medical teachers to identify which station needs to be revised or discarded.

Another analysis that can increase the homogeneity of an OSCE exam is the use of Pearson's correlation which can be used to find a correlation between mean station scores and the mean total OSCE score. As each station is contributing to OSCE homogeneity, those stations that do not correlate with the OSCE exam as a whole can be revised or discarded. The homogeneity and heterogeneity of a test or an OSCE exam is an important issue that is further discussed in the next section.

Homogeneity and heterogeneity of the items

If the items of a test measure a single feature, the test is termed homogenous. In other words, homogeneity is the extent to which a test taps a single domain and does not include items that measure other abilities. For example, a test of cardiovascular physiology should assess knowledge of the cardiovascular system, not all medical physiology. It should be noted that items on a test should come from a random sample of the item pool and measure a single domain. These items should also correlate with each other to varying degrees (Streiner 2003).

In contrast to test homogeneity, the items of a heterogeneous test tap different domains or attributes. In the above example the items on the cardiovascular system tap one area while the items on medical physiology not only measure the cardiovascular system but also measure renal, lung, gastrointestinal systems and so forth. Those who receive the same score on a multiple-choice homogenous test have a similar knowledge in the area tested. On the other hand, those who receive the same score on a multiple-choice heterogeneous test may have different knowledge in the areas tested (Cohen & Swerdlik 2010). This simply illustrates that test scores that come from a heterogeneous test are more ambiguous than a homogenous test. Imagine that in a heterogeneous medical physiology test, John and Sarah both receive a score of 30. One cannot conclude that knowledge or performance of both on the test was equal. The score of 30 can be obtained through a variety of combinations. John may have correctly answered 10 cardiovascular physiology items, 10 sensory physiology items, 10 respiratory physiology items and none on neurophysiology or gastroenterology. Sarah by contrast, may have correctly answered 5 sensory physiology items, 10 respiratory physiology items and 15 neurophysiology items and none on cardiovascular physiology. If more specific assessment data is required it is better to develop several homogenous tests in which each test measures a single domain.

The homogeneity of a test is also an indicator of construct validity as it ensures that all the questions on the test measure the same construct or trait. It should be noted that test designers should determine the validity of a test or an OSCE exam before an examination in order to assess the degree to which the test accurately reflects the specific trait that the test designer is attempting to measure.

The standard Error of measurement (SEM)

One final useful concept concerned with post-exam analysis is the standard error of measurement (SEM). The SEM provides an estimate of the amount of error inherent in an individual's test score (Cohen & Swerdlik 2010). This estimation helps assessors to determine the discrepancies between an individual's observed score on the test and his/her true score. There is a link between the test reliability estimate and the SEM. The larger the test reliability estimate, the lower the SEM. If the estimate of the reliability of a test and its standard deviation are determined, the SEM is calculated by the following below:

$$\mathrm{SEM} = \mathrm{SD}\sqrt{1 - \mathrm{r}}$$

Where SEM is equal to the standard error of measurement, SD is equal to the standard deviation of test scores by a group of students and r is equal to the reliability coefficient of the test. Assuming a medical student achieved a score of 50 (out of 100) on a test. If the test had a standard deviation and a reliability coefficient (e.g. split—half reliability) of 10 and 0.74, respectively, then the SEM is 5 ($\mathrm{SEM} = 10\ \sqrt{1 - 0.74}) = 5$).

Before interpreting the value of the SEM it is helpful to know that in a normal distribution roughly 68% of the values lie within ± 1 standard deviation of the mean, 95% of the values lie within ± 2 standard deviation of the mean and 99.75% of the values lie within ± 3 standard deviation of the mean. Assuming the distribution of cardiovascular test scores is normal we can now estimate the true score for the student as shown below.

We can be 68% confident that his true score lies within $50 \pm 1_{\mathrm{SEM}}$ (or between 45 and 55), 95% confident that the true score lies within $50 \pm 2_{\mathrm{SEM}}$ (or 40 and 60) and 99% confident that the true score lies within $50 \pm 3_{\mathrm{SEM}}$ (or 35 and 65).

The SEM also aids in decision making about a students' performance on the test. If standard setters, in the above test, set a cut score for failing of 50 and if assessors want to be 68% confident of their decision, the SEM indicates that the student's true score, lies between 45 and 55. This means that if the student was to take the test again, his/her score might be less or more than the cut score (between 45 and 55). This indicates that other student activities need to be taken into account when deciding whether or not the student should pass the test.

Qualitative item analysis

Finally it is worthwhile being aware that there are non-statistical, qualitative methods, of ensuring the quality of objective test items. Test constructors have had a long-standing

interest in the way students make sense of their experiences on tests (Mosier 1947; Fiske 1967). Qualitative methods can be employed to explore the meanings students attach to their experience following a test and how they make sense of that test. Researchers can immerse themselves in the natural setting of students who have taken a particular test. Exploring the interaction between the test constructor and the student provides the opportunity for a deep understanding of the items under investigation.

Qualitative methods utilise techniques for generating and analysing data which is grounded in the voice of students rather than psychometric-statistical inferences. In other words, the units of analysis are the words of students rather than their numerical scores.

'The student voice' can be gleaned from different verbal sources such as group interviews, face to face interviews or observations. The purpose of the interview is to explore students' subjective understanding of their test-taking experience. Qualitative test constructors seek to uncover how individual test items work. Test developers usually construct an interview schedule containing open and closed questions to uncover potential areas of exploration by means of qualitative analysis. The potential areas that may be contained in the interview schedule are: cultural awareness, test validity, test administration, test environment, test fairness, test language, item guessing, student preparation, student's comfort during the test, test length, test time and overall impression of the student (Cohen & Swerdlik 2010).

'Think aloud' test administration is an observation qualitative research instrument to uncover student's responses to each item or skill during the administration of a test. In this approach, students are asked to take part in a test and then express whatever they are feeling and thinking when they are responding to each item or skill. Examiners make objective notes of students' utterances or audio-record them, without interruption, during the test. Transcriptions and analysis of the materials is carried out using qualitative research methods. Such verbalisations by students may help examiners to better understand how students interpret an item, as well as why and how they are misinterpreting an item (Cohen & Swerdlik 2010).

It should be noted, however, that students' scores may influence their responses to the questions during interview. Those who have received good scores may respond positively and those who have received poor scores may criticise test developers. The interpretation of qualitative data should take all student experiences into consideration. Based on these interpretations, examiners or test developers can revise, reword or discard an item.

Summary

This Guide has explained the central importance of measurement and evaluation and inferential foundations of examination questions in medical education. Medical educators have three key roles in facilitating student engagement in learning. First, they need to make a decision about learning objectives which focuses on what medical students need to do or know. Second, medical educators need to implement and teach the target subject matter in health care settings or the classroom using educational management and leadership techniques and pedagogical methods. Finally, medical educators need to measure and evaluate how much of the material has been mastered by a particular achievement test. This test is usually considered as a criterion for student achievement in a particular subject. Consequently, medical educators need to construct valid and reliable tests in order to ensure that examination questions elicit evidence that is appropriate to the intended purpose. To this end, the item-difficulty index and the item- discrimination need to be calculated. A question is considered easy if it is answered correctly by the majority of students (more than 60%), and is considered hard if it is answered correctly by less than 30% of the students. The item-discrimination index is analysed using the point-biserial correlation coefficient and the *t*-test procedure. A large positive R_{pbi} is an indication of a good question while a low positive or a negative R_{pbi} is an indication of a bad question. The t-test is another statistical procedure for determining the item-discrimination index. If there is no significant difference between the mean score of students who answered the question correctly and the mean score of students who answered the question incorrectly, the question is not differentiating strong students from weak students. This suggests that the question should be removed or revised for next examination. SPSS facilitates the analysis of the item analysis data.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

Notes on contributors

MOHSEN TAVAKOL, PhD, MClinEd is an independent medical education consultant. He is an Assistant Professor of Medical education, Tehran University of Medical Science. His main interests are in medical education assessment, psychometric analysis, quantitative and qualitative research methods and communication skills. He is an editor of the on-line journal International Journal of Medical Education.

REG DENNICK, PhD, MEd, FHEA is a Professor of medical education in the University of Nottingham. His main interests are in medical education teaching and research, problem-based learning, staff training and curriculum development.

Note

1. A full version of Post Examination Analysis of Objective Tests: AMEE Guide no 54, by Tavakol & Dennick can be purchased in hard copy or in PDF through the AMEE office (www.amee.org). The full Guide provides, in addition, a step by step description of the methods of analysis using SPSS as well as a recommended reading list.

References

Anastasi A, Urbin S. 1997. Psychological testing. Upper Sadder River, New Jersey: Prentice Hall.

Anderson L, Krathwohl D. 2000. A Taxonomy for learning, teaching, and assessing. Boston: Longman.

Bandaranayake R. 2008. Setting and maintaining standards in multiple choice examinations: AMEE Guide No.37'. Med Teach 30(9):836–845.

Bergstrom B, Lunz M. 2008. CAT for certification and licensure. In: Drasgow F, Olson-buchanan J, editors. Innovations in computerized assessment. Mahwah, NJ: Lawrence Erlbaum Associates.

Biggs J, Tang C. 2007. Teaching for quality learning at university. Maidenhead, England: Open University Press.

Bloom B. 1956. Taxonomy of educational objectives. Boston, MA: Allyn and Bacon.

Case S, Swanson D. 2010. Constructing written test questions for the basic and clinical sciences (3rd edition). National Board of Medical Examiners. Available to download from: http://www.nbme.org/publications/item-writing-manual.html

Cohen R, Swerdlik M. 2010. Psychological testing and assessment. Burr Ridge, IL: McGraw-Hill.

Cortina J. 1993. What is coefficient alpha? an examination of theory and applications. J Appl Psychol 78:98–104.

Crocker L, Algina J. 2008. Introduction to classical and modern test theory. Mason, Ohio: Cengage Learning.

Cusimano M. 1996. Standard setting in medical education. Acad Med 71(10):S112–S120.

Davis F. 1952. Item analysis in relation to educational and psychological testing.. Psychol bull 49:97–121.

Dechamplain AF. 2010. A primer on classical test theory and item response theory for assessments in medical education. Med Educ 44:109–117.

Dreyfus H, Dreyfus S. 2000. Mind over Machine. New York: Simon & Schuster.

Fiske D. 1967. The subjects react to tests. Am Psychol 22:287–296.

General Medical Council. 2003. Tomorrow's Doctors. Recommendations on undergraduate medical education. UK: GMC.

Gilbert S. 1996. Principles of educational and psychological measurement and evaluation. Belmont, CA: Wadsworth Publishing.

Glaser R. 1963. Instructional technology and the measurement of learning outcomes. Am Psychol 18:510 GMC 522.

Green B, Bock R, Humphreys L, Linn R, Reckase M. 1984. Teaching guidelines for assessing computerised adaptive tests. J Edu Meas 21:347–360.

Gregory R. 2007. Psychological testing: history, principles and applications. Boston: Pearson Education, Inc.

Henson R. 2001. Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. Meas Eval Couns Dev 34:177–189.

Holmbow E, Hawkins R. 2008. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby Elsevier.

Hopkins K. 1998. Educational and psychological measurement and evaluation. Boston: Allyn and Bacon.

Ideal Consortium. 2010. International database for enhanced assessment and learning. Hong Kong: Ideal Consortium. http://www.hkwebmed.org/idealweb/homeindex.html

Kaplan R. 2008. Psychological testing: principles, applications, and issues. Belmont, CA: Wadsworth.

Kelley T. 1939. The selection of upper and lower groups for the validation of test items. J Edu Psychol 30:17–24.

Leeds Institute for Medical Education. 2010. Universities medical assessment partnership. Http://www.leeds.ac.uk/medicine/meu/cpd/umap.html

Mcalpine M. 2002. A summary of methods of item analysis: Bluepaper number 2. Computer assisted assessment centre. University of Glasgow: Robert Clark Centre for Technological Education.

Morrow J, Jackson A, Disch J, Mood D. 2006. Measurement and evaluation in human performance. United States: Human Kinetics.

Mosier C. 1947. A critical examination of the concepts of face validity. Edu Psychol Meas 7:191–206.

Norcini J. 2003. Setting standards on educational tests. Med Educ 37:464–9.

Norcini J, Guille R. 2002. Combining tests and setting standards. In: Norman G, van der Vleuten C, Newble D, editors. International handbook of research in medical education. Dordrecht: Kluwer Academic Publisher.

Osterlind SJ. 1998. Constructing test items: multiple-choice, constructed-response, performance and other formats. Boston: Kluwer Academic Publishers.

Rust J, Golombok S. 2009. Modern psychometrics. London: Rutledge.

Schmitt N. 1996. Use and abuse of coefficient alpha. Psychol assessment 8:350–353.

Shultz K, Whitney D. 2005. Measurement theory in action. London: Sage Publications.

Simpson J. 1966. The classification of educational objectives: psychomotor domain. Office of education project no. 5-85-104. Urbana, IL: University of Illinois.

SPSS Inc. 2009. SPSS Base 17.0 for Windows User's Guide. Chicago, IL: SPSS Inc.

Streiner D. 2003. Being inconsistent about consistency: When coefficient alpha does and does not matter. J Pers Assess 80:217–22.

The Scottish Deans' Medical Curriculum Group. 2007. http://www.scottishdoctor.org/

Traub RE, Rowley GL. 1991. NCME instructional module: Understanding reliability. Educ Measur Issue Pract 10:171–9.

Tyler R. 1949. Basic principles of curriculum and instruction. Chicago: University of Chicago Press.

Weiss D, Vale C. 1987. Computerized adaptive testing for measuring abilities and others psychological variables. In: Butcher J, editor. Computerized psychological assessment: a Practitioner's guide. New York: Basic Books.

World Federation for Medical Education. 2003. Basic medical education. WFME global standards for Quality Improvement. WFME, Copenhagen 2003. WFME website: http://www.wfme.org