A primer on classical test theory and item response theory for assessments in medical education

André F De Champlain

CONTEXT A test score is a number which purportedly reflects a candidate's proficiency in some clearly defined knowledge or skill domain. A test theory model is necessary to help us better understand the relationship that exists between the observed (or actual) score on an examination and the underlying proficiency in the domain, which is generally unobserved. Common test theory models include classical test theory (CTT) and item response theory (IRT). The widespread use of IRT models over the past several decades attests to their importance in the development and analysis of assessments in medical education. Item response theory models are used for a host of purposes, including item analysis, test form assembly and equating. Although helpful in many circumstances, IRT models make fairly strong assumptions and are mathematically much more complex than CTT models. Consequently, there are instances in which it might be more appropriate to use CTT, especially when common assumptions of IRT cannot be readily met, or in more local settings, such as those that may characterise many medical school examinations.

OBJECTIVES The objective of this paper is to provide an overview of both CTT and IRT to the practitioner involved in the development and scoring of medical education assessments.

METHODS The tenets of CCT and IRT are initially described. Then, main uses of both models in test development and psychometric activities are illustrated via several practical examples. Finally, general recommendations pertaining to the use of each model in practice are outlined.

DISCUSSION Classical test theory and IRT are widely used to address measurement-related issues that arise from commonly used assessments in medical education, including multiple-choice examinations, objective structured clinical examinations, ward ratings and workplace evaluations. The present paper provides an introduction to these models and how they can be applied to answer common assessment questions.

*Medical Education 2010: **44**: 109–117*
doi:10.1111/j.1365-2923.2009.03425.x

National Board of Medical Examiners, Philadelphia, Pennsylvania, USA

Correspondence: André F De Champlain, National Board of Medical Examiners, 3750 Market Street, Philadelphia, Pennsylvania 19104, USA. Tel: 00 1 215 590 9565; Fax: 00 1 215 590 9449; E-mail: adechamplain@nbme.org

INTRODUCTION

Examinations are part and parcel of every phase of a medical professional's training. They provide important information about a student's progress through his or her education. The score on a paediatrics examination obtained following a clerkship in that clinical area, for example, can be a key indicator of the extent to which the medical student has mastered knowledge of that domain. However, any examination, by virtue of practical constraints (e.g. testing time), contains a sample of all test items that comprise the domain, representing the theoretically infinite pool of items that targets content and skill areas of interest. Our paediatrics examination might include 100 items administered over 2 hours, although it is not difficult to envisage another similar test that might include a different set of 100 items that are equally useful. Measurement occurs when a number (a test score), assumed to reflect performance in the domain, is assigned. However, as illustrated above, it is possible to imagine a scenario in which two satisfactory examination forms yield different scores for the same candidates. This example highlights the facts that no examination is perfect and that all scores contain measurement error unrelated to the targeted domains. Test theory models are necessary to help us better understand the measurement process and how it is impacted by sources of error. Additionally, as abilities are generally unobserved, a test theory model is needed to better explain the relationship that exists between actual test scores and estimated performance in the domain. By unobserved, we mean that the candidate's true ability is inferred from his or her score on an examination. For example, a given candidate's score on our 100-item paediatrics examination is a reflection of his or her true knowledge of that content area; that is, the latter measure cannot be observed (or scored), but, rather, is estimated based on the sample of items that comprises the examination.

Two main test theory models have been proposed for creating and evaluating examinations: classical test theory (CTT) and item response theory (IRT). The goal of this paper is to provide a primer that will enable the practitioner to better comprehend the foundations of both approaches. Both CTT and IRT have been used in the assessment of medical students, from the early days of undergraduate training to postgraduate education, to help in developing examinations that are well targeted to candidate abilities in terms of their difficulty.[1] Similarly, these two frameworks are heavily used in all phases of activity in large-scale medical certification and licensure programmes, from test development efforts to scoring and reporting tasks.[2] The secondary aim of this paper is therefore to help the practitioner determine instances in which one approach might be preferable over another, or whether both models might provide useful information in light of sample size and other practical issues that may characterise a particular assessment.

OVERVIEW OF CLASSICAL TEST THEORY

A general framework

The central tenet of classical test theory[3–5] is that the score that a candidate obtains on a given examination, which is symbolised by (X), can be decomposed into the person's true score (T) and a random error component (E):

$$X = T + E \quad (1)$$

The candidate's true score, T, is defined as the expected value of the observed score over an infinite number of repeat administrations with the same examination. A true score can be thought of as the score that would be obtained if the examination was perfectly measuring the ability of interest (i.e. with no measurement error). If our 100-item paediatrics examination could perfectly measure our candidate's knowledge of that clinical science, then the observed and true scores would be equal. As this never happens in practice, it becomes important to assess the extent to which an actual test score (computed from a sample of items comprising the examination) reflects true knowledge of the domain(s) presumably being targeted by the test. A reliability coefficient can provide us with an estimate of the level of concordance between observed and true scores.

Classical test theory reliability

A reliability coefficient provides us with an estimate of the level of precision with which a score on an examination reflects the candidate's true score.[6,7] Mathematically, the reliability coefficient is expressed as the proportion of observed score variance (σ^2_X) shared or attributed to true score variance (σ^2_T). The reliability coefficient can also be viewed as the correlation between scores on two truly parallel forms of a test. Parallel forms can be defined as examinations that measure the same content, and on which candidates have the same true score, with equal errors of measurement across forms.[8] Of course, the

reliability coefficient is a theoretical concept which needs to be estimated using observed data. As reliability focuses on accuracy of measurement, it becomes necessary for the user to define the facet over which scores are intended to be reproducible. For example, in a ward evaluation of clinical performance, examination scores are typically derived from expert ratings. In this instance, the examination director might be particularly interested in assessing precision of measurement across raters (i.e. inter-rater reliability). In this setting, reliability might be estimated by computing the percentage of agreement between raters with regard to judgements, with or without correction for chance agreement. Alternatively, more sophisticated models, such as generalisability theory, allow the user to estimate reliability as a function not only of the judges participating in an evaluation, but also of the clinical tasks, setting and any other facet that might impact the precision of measurement. Applications of these models to medical education can be found elsewhere.[9] Conversely, with multiple-choice question (MCQ) examinations, methods requiring a single test administration, in particular internal consistency indices, are more commonly computed. In particular, Cronbach's coefficient alpha (α) is popular and provides an estimate of the proportion of observed score variance in the candidate's performance across items within a test that is attributable to true score variance.[10]

Within the context of mastery testing, it is often of greater importance to assess decision consistency; that is, the extent to which we are accurately classifying candidates as masters and non-masters. With this type of examination, precision at the cut-score, rather than along the entire score scale, is of greater importance. Several indices have been proposed for estimating mastery reliability and can be found elsewhere.[11]

Standard error of measurement

At the individual test score level, we are often more interested in computing an expected measure of error. The standard error of measurement (SEM) provides such an estimate and is given by:

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}} \qquad (2)$$

where

σ_X = the total test score standard deviation (SD), and
ρ_{XX} = the reliability coefficient estimate.

If we could test a candidate an infinite number of times with the same examination (assuming no learning occurred between administrations!), the mean of all scores would be equal to the true score, whereas the SD would correspond to the SEM. As with any standard error (SE), the SEM is typically used when establishing a confidence interval within which we expect a candidate's true scores to fall. Readers interested in obtaining more information on these computational details should consult another source.[12]

Item-level statistics

In medical schools or other educational settings, one of the central purposes of test construction is to develop an examination (of minimal length) that will allow us to accurately measure our candidates' abilities in the domains of importance. Given the practical constraints, which may include testing time, faculty availability and funding, how can we develop the best assessment for our purposes? In order to attain this goal, for either MCQ examinations or performance assessments, it is often necessary to field-test a larger number of items or tasks to arrive at a final set that will actually comprise the examination. The latter process is referred to as item analysis and typically entails computing item difficulty and discrimination indices.

With dichotomously scored items (i.e. items scored as incorrect or correct), item difficulty is generally denoted as a *P*-value, reflecting the proportion of candidates who correctly answer a given test item. Lower *P*-values are indicative of more difficult items because fewer candidates give a correct response. Conversely, higher *P*-values suggest easier items as they are correctly answered by a higher proportion of the cohort. With performance assessments, such as an objective structured clinical examination (OSCE), the mean rating scale value can also be computed for each station as an indicator of difficulty.

It is also useful to assess whether items discriminate between candidates of varying ability. In an MCQ examination, if the test item is functioning as intended, we would expect a higher proportion of more able than lower-ability candidates to correctly answer a given item; that is, we would expect the item to *discriminate* between low- and high-ability candidates. Common item discrimination indices are correlational in nature and include the point-biserial and biserial coefficients. Both coefficients provide an estimate of the correlation between the response on a test item (0 or 1) and a criterion measure, usually the total test score. Similarly, for performance assessments, polyserial correlation coefficients can be computed between ratings on a given station or clinical situation and the total examination score.

APPLICATIONS OF CTT IN MEDICAL EDUCATION

The main advantage of CTT is that it is based on relatively weak assumptions that are easy to meet with real data and modest sample sizes. These models are simple to use and require little mathematical knowledge on the part of the user. In most medical education settings, where the aim is to develop assessments that will be used locally, with little or no intention to generalise beyond that setting, CTT is very useful for assessing the difficulty and discrimination of items, as well as the precision with which scores are measured by an examination. How might a faculty member use CTT to develop a local examination?

Once content specifications and other test constraints (number of items, testing time, etc.) have been outlined in a blueprint and other guiding documents, the first data analytic step would involve summing up item responses to create a total test score (sum of item 1s and 0s). Following this, it is generally of interest to compute item difficulty and discrimination indices to gain a sense of how items perform in light of the objectives of the assessment. The targeted level of difficulty will largely depend on the purpose of the examination. If the goal of the examination is to be able to rank-order candidates in a clerkship with the highest level of precision, as might be the case in many local assessments, then it may be preferable to retain items that have difficulty values in the 0.3–0.7 range (with a mean of 0.5). Items with a mean *P*-value of 0.5 yield the highest item variance because the latter value is equal to p^*q, where q is the proportion of candidates who *incorrectly* answered an item. Thus, items with *P*-values of 0.5 allow us to separate the weakest from the most able candidates with the highest level of precision or reliability (item variance = 0.5×0.5 or 0.25). However, in a criterion-referenced testing situation, where a decision is made based on meeting a pre-specified cut-score, such as for passing a course or for promotion, it might make more sense to select items with item difficulty values near the cut-score as they provide information in the vicinity of the score distribution that is of greatest importance.

Similarly, item discrimination values can be computed to help identify potential keying errors or items that may not be functioning as intended (e.g. negatively discriminating items or items that weaker students tend to answer correctly in a higher proportion than more able candidates). As a rough rule of thumb, items that have low point-biserial values (e.g. ≤ 0.2) could be flagged for review by content experts to determine whether there is a substantive reason that supports retaining or excluding them from the scoring process.

At the total test score level, a reliability coefficient should also be computed to assess the level of precision with which our candidate abilities are being measured by a given examination form. For an MCQ examination, Cronbach's coefficient α should be calculated as an estimate of score reliability that can help the user assess whether candidates can be accurately rank-ordered from least to most able. What constitutes 'acceptable' reliability? The answer to this question depends largely on the type of assessment, the stakes, and the intended interpretation of the test score. In a high-stakes situation with MCQs, it is generally advised to have a reliability estimate ≥ 0.90. However, with performance assessments, such as OSCEs, where testing time severely curtails the number of stations that can be administered, a reliability estimate of 0.70 might be more realistic. Similarly, in a lower-stakes formative assessment context, reliability estimates of ≥ 0.70 might be acceptable. Regardless of the context and assessment, it is critical to also compute the SEM to aid the user in gauging the accuracy with which scores are being estimated and also to temper any interpretation of scores that may not be justified given measurement imprecision. For example, it would be unadvisable to make high-stakes decisions with examination scores that have a reliability of 0.70, given that an observed candidate's score could be a fairly poor reflection of his or her true score in that particular domain.

LIMITATIONS OF CTT

Although valuable, the use of CTT to assemble examinations and to analyse data is not without its limitations. First and foremost, all CTT-based statistics are sample-dependent. For example, the *P*-value associated with an item reflects not only the difficulty of the content matter targeted, but also the ability level of candidates answering the question. A *P*-value of 0.60 for a given test item will not represent the same level of difficulty if it is based on a very weak cohort that it will when based on a very able group of candidates. Consequently, CTT is useful for form assembly only in instances where groups of candidates are comparable in ability. In addition, CTT does not provide an easy mechanism by which to target an examination at a certain ability level, which is an important consideration for

mastery tests where the goal is to maximise reliability or precision of measurement at one point on the score scale, rather than for the entire range of scores. Finally, CTT assumes that measurement error is identical for all scores. In practice, however, we know that scores located in the (sparser) tails of the distribution are not estimated as accurately as those located in the middle region as a result of the paucity of information contained in that region (i.e. score estimates are poor when based on small numbers).

In certain instances, for example with a graduation examination, where it might be important to track scores and item difficulties across years, forms and cohorts, CTT might not be appropriate for the reasons mentioned above. When certain conditions are met, IRT can be useful in addressing many of the shortcomings of CTT.

OVERVIEW OF ITEM RESPONSE THEORY

Item response theory encompasses a family of non-linear models that provide an estimate of the probability of a correct response on a test item as a function of the characteristics of the item (e.g. difficulty, discrimination) and the ability level of test takers on the trait presumably being targeted by the test form.[13] All IRT models attempt to explain observed (actual) item performance as a function of an underlying ability (unobserved) or latent trait. Two common IRT models for dichotomously-scored MCQ examinations are the one-parameter logistic (1-PL)/Rasch and the two-parameter logistic (2-PL) models.

The 1-PL IRT model is given by:

$$P_i(x_i = 1|b_i, \theta_j) = [1 + e^{-D(\theta_j - b_i)}]^{-1} \qquad (3)$$

where

P_i (x_i = 1) = the probability of correctly answering item i, given:

b_i = the difficulty of item i, and θ_j = the ability level of candidate j.

Additionally, the following two constants are included in the model: e, which is the base of the natural logarithm scale ($\sim$ 2.7178), and D ($\sim$ 1.7), which is used to approximate a normal ogive model. With this model, we can estimate the probability that a candidate will correctly answer a test question given one item parameter or value (i.e. item difficulty), as well as from an estimated ability on the entire examination. We can also represent this model graphically in an item characteristic curve (ICC). Sample ICCs for two test items based on a 1-PL model are shown in Figure 1. The x-axis of the ICC corresponds to ability estimate values. Higher theta (θ) values are associated with more able candidates. The probability of a correct response to the item is shown along the y-axis (ranging from 0 to 1). The more able a candidate is, the more likely he or she will correctly answer the item. Item difficulty corresponds to the ability estimate (θ_j) that is associated with a probability of 0.5 of a correct response. In IRT, higher positive b-values reflect more difficult items, whereas lower negative values are indicative of easier items. In our example, item 2 (b = + 0.5) is more difficult than item 1 (b = − 0.5).

The 2-PL model can be written as:

$$P_i(x_i = 1|b_i, a_i, \theta_j) = [1 + e^{-Da_i(\theta_j - b_i)}]^{-1} \qquad (4)$$

where P_i (x_i = 1), b_i and θ_j have been previously defined, and where a_i corresponds to the discrimination parameter value for item i. Sample ICCs for two items analysed with a 2-PL model are shown in Figure 2. Both items in this example are of equal difficulty (0.0). However, item 2 has a higher discrimination parameter value (1.0) than item 1 (0.5), as evidenced by the steeper ICC. Thus, item discrimination defines the slope or steepness of the ICC. If our two sample items were part of an assessment in paediatrics, item 2 would be better able to discriminate between candidates who scored below and above an ability value of 0. Item 2 would be particularly well suited if it was part of a mastery test where the cut-score corresponded to a θ-value near 0.

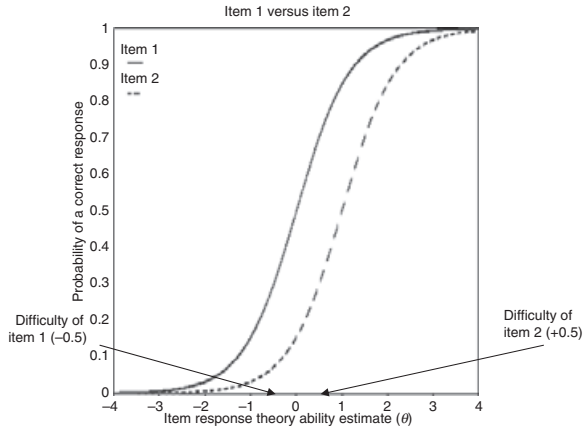

Figure 1 Sample 1-PL (one-parameter logistic) item response theory item characteristic curves for two items

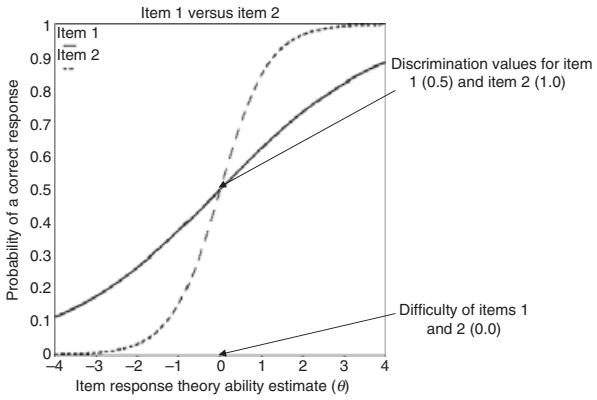

Figure 2 Sample 2-PL (two-parameter logistic) item response theory item characteristic curves for two items

WHICH IRT MODEL SHOULD BE USED IN MEDICAL EDUCATION ASSESSMENTS?

The 1-PL model is particularly useful in the majority of medical education settings, largely because it requires fewer examinees than the 2-PL model. This is because we are estimating a smaller number of parameters in a 1-PL framework: that is, item difficulties only. Sample sizes that have been proposed in the literature are 500 examinees for the 2-PL model and 200 for the 1-PL model.[14,15] Thus, the use of the 1-PL model is feasible with larger class sizes, whereas there are few instances where 2-PL sample size requirements might be met in a medical education context. Some of the more complex IRT models are used in large-scale achievement testing programmes where sample sizes exceed tens of thousands per annum. Most medical certification and licensure examination programmes use the Rasch or 1-PL model for a number of reasons. Firstly, in the latter framework, there is a one-to-one correspondence between each raw score and ability estimate. Two candidates who correctly answer 75/100 items on a graduation examination will have the same IRT ability estimate. For example, in a given IRT calibration, a raw score of x will always correspond to the same θ-value of y. A calibration is the process by which we obtain estimates of item difficulties and candidate abilities. However, the 1-PL model makes stronger requirements than alternative models, namely, that discrimination values do not significantly vary across items.

Unlike the 1-PL model, a given raw score can correspond to multiple ability estimates in a 2-PL framework, as the latter depend on both the difficulty and discrimination values of the items encountered. For example, on the same graduation examination, two candidates who correctly answered 60/100 items may have a different ability estimate, depending on which items were correctly answered. However, the 2-PL model provides an estimate of discrimination for each item and might be more appropriate with datasets that display wide variation in this item characteristic, assuming that appropriate sample sizes are available.

COMMON PROPERTIES OF IRT MODELS

Unlike CTT models, IRT models are particularly appealing as a result of their parameter invariance properties.[16] If the assumptions of common IRT models are met with the dataset, then item parameter estimates are independent of the particular sample of examinees drawn from the population of examinees for whom the test is intended (*item parameter invariance property*). Further, an examinee's estimated ability is not dependent upon the particular sample of test items chosen from the calibrated pool of items (*ability parameter invariance property*). Finally, IRT models also provide an estimate of measurement error at each point along the ability (θ) scale.

The first two properties enable the comparison and tracking of candidates who may have completed different test forms, as long as both groups answer a common subset of test questions. It is therefore possible to track both the difficulty level of graduation examinations in one discipline across years and to compare the ability level of different classes, regardless of the test form completed. It is not possible to do this within a CTT framework, given the sample dependence issues previously outlined. The third property is especially appealing with a mastery test, where it is important to maximise precision of measurement at a specific point on the score scale (i.e. at the cut-score). Item response theory provides us with a tailored estimate of measurement error for all abilities, including the one that corresponds to the cut-score. This is another attractive feature over CTT, which only yields one general measurement estimate for all scores. Of course, the usefulness of these features depends on how well the model actually fits the dataset.

IRT MODEL FIT REQUIREMENTS

Item response theory holds several attractive advantages over CTT, particularly as it is able to estimate and compare student abilities regardless of the test forms completed. However, to take full advantage of these models, several assumptions need to be met. Firstly, common IRT models assume that a single underlying ability accounts for performance on the examination. This is referred to as the assumption of

unidimensionality. However, past research has shown that IRT models are robust to some level of departure from this assumption, if the composite ability is stable across test forms.[16] Nonetheless, in instances where an examination measures different domains, which can vary across test forms, common IRT models should not be used to estimate item difficulties and candidate abilities. Item response theory might not be particularly well suited to OSCEs and other complex performance assessments that often target several clinical skills. The most straightforward way to test the assumption of unidimensionality empirically is to use factor analysis, specifically by assessing the fit of a one-factor model to a dataset.[17]

Related to this first assumption is local independence (LI), which is met when the conditional probability of a correct response to an item (conditional on a candidate's ability) is unrelated to the conditional probability of correct responses to other items in the test. Plainly stated, common IRT models assume that an answer to one item is unrelated to the answer to any other item on the test. A final assumption is that tests are non-speeded. Speeded tests introduce additional dimensions that are unrelated to the ability targeted by the examination. With a computer-based examination, it is possible to assess speededness by computing the number of seconds spent on each test question to see whether patterns differ for items that appear towards the end of an examination, which are more likely to be affected by the candidate running out of testing time. In addition, for the 1-PL model only, it is assumed that discrimination parameters are equal across items. Plotting point-biserial correlations for each item provides a quick check on the extent to which this assumption holds and whether we are justified in using a 1-PL IRT model or not.

Finally, common IRT software packages provide overall model-fit statistics that can also be useful. Most of these statistics provide an indication of how closely the model-estimated probability of a correct response (as shown in Eqns 3 and 4) coincides with the actual proportion of correct responses at each ability level. Readers interested in obtaining more information on these statistics are referred to other sources.[18,19]

APPLICATIONS OF IRT IN MEDICAL EDUCATION

Assessing reliability in IRT: item and test information functions

In an IRT framework, information functions provide a graphical representation of the precision of measurement at each ability level for either an individual test item (item information function [IIF]) or the entire test (test information function [TIF]). The greater the information present at a given ability level, the more precise or reliable the measurement will be at that θ-value. Item information is additive; IIFs are added to obtain a TIF. For 1-PL and 2-PL IRT models, information is always maximised at the ability estimate which corresponds to the difficulty parameter value.

Similarly, the concept of the SE of the ability estimate (SE_θ) is analogous to that of the SEM in CTT. However, unlike the traditional SEM, a separate IRT-based SE is computed for each θ-value. It is therefore possible to assess how reliably we are measuring each score in the distribution, including the passing standard in examinations on which mastery decisions are made. Computational details for both information and the SE of the ability estimate can be found elsewhere.[13]

What should a TIF look like with examinations routinely used in medical education? The answer to this question depends on the intended use of test scores. Figure 3 presents sample TIFs for a medical certification examination (dashed line) and a selection examination (solid line). The Medical College Admission Test and the Biomedical Admissions Test are examples of selection examinations. With selection examinations, it is important to measure a broad range of abilities with a similar level of precision or reliability out of fairness to candidates who are

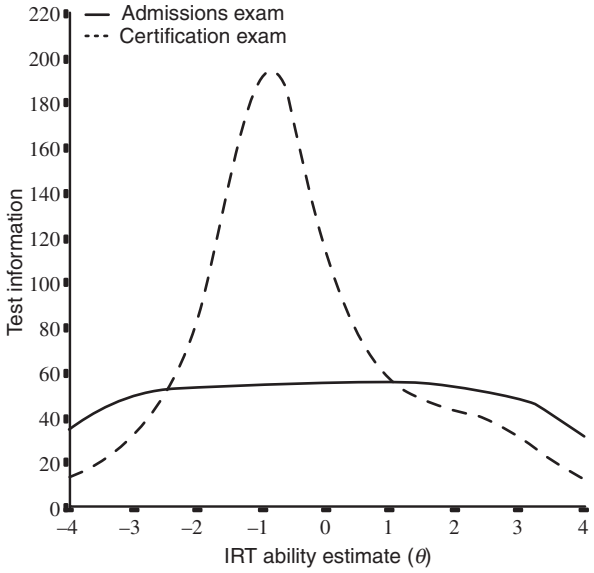

Figure 3 Sample test information functions

applying for entry into medical school. Moreover, a university office needs assurances that the information it is factoring into admissions decisions is comparable across candidates. However, in the case of a medical licensure or certification examination, reliability or information needs to be maximised at the cut-score value because this is where decision accuracy needs to be at its highest point. Consequently, the dashed TIF shown in Figure 3 is appropriate in the instance where the cut-score would correspond to an ability estimate of −1.

Test assembly and banking

Item response theory models are routinely employed at the test assembly stage. Typically, a target TIF is specified and items are selected to produce the desired curve. After each item is added, the TIF is calculated; this process continues until the target function is attained to a satisfactory degree. This approach is used heavily in medical licensure and certification examination programmes where security concerns make it necessary to administer multiple forms of the same examination at any one time. In this instance, it is critical to administer test forms that are comparable in regard not only to content, but also to precision of measurement.

Within test development efforts, the invariance properties of common IRT models also make it possible to bank item statistics (computed across different cohorts of candidates) in a common repository for future assembly and use. This is routinely done with medical licensure examinations, such as those that comprise the United States Medical Licensing Examination (USMLE®). Item banks can also be developed at the medical school level, especially for summative assessments with larger class sizes such as end-of-course examinations. It might be feasible in those contexts to compute 1-PL statistics across forms from year to year to better inform future test development activities.

Computer-based testing

Computer-based tests (CBTs) are now part and parcel of many large-scale medical assessments as they hold several advantages over traditional paper-and-pencil tests for both the candidate and the testing agency. From the candidate's perspective, the CBT offers greater flexibility with respect to the schedule and the test administration site. From the testing agency's perspective, the CBT potentially augments the validity of score- and decision-based interpretations by affording greater control over security while still meeting strict reliability requirements.[20] Item response theory models are heavily used in all phases of CBT activities, from test assembly to scoring and reporting activities.

With computer-adaptive tests (CATs), where an algorithm tailors a unique examination to each candidate to provide maximum precision of measurement at his or her specific ability level with the fewest number of items possible, IRT models are central. In one CAT scenario, the probability of a correct response is computed for an initial item of average difficulty using an IRT model. The candidate's answer to this item determines whether an easier or more difficult question is subsequently administered. That is, the examination 'adapts' to the candidate's ability until precision of measurement can no longer be improved. Although CATs are attractive in that they can yield precise measurements of abilities with shorter test lengths, very large item banks are necessary to sustain such programmes. Readers interested in obtaining more information on this testing modality should consult other sources.[20]

USING IRT WITH PERFORMANCE ASSESSMENTS

Most of the examples discussed throughout this paper relate to MCQ examinations. However, many assessments used in medical education incorporate rating scales, such as in OSCEs and clerkship ratings. Variants of the IRT models outlined in this paper can be particularly useful for performance-based assessments. In particular, adaptations of the Rasch or 1-PL model can be applied to estimate not only the difficulty of tasks (e.g. an OSCE station or clinical scenario), but also the stringency level of raters and any other facet that may be impacting measurement precision. However, these models are still subject to the same assumptions of unidimensionality and local independence. Readers interested in obtaining an overview of common models and applications should refer to other sources.[21]

SOFTWARE

Several commercial IRT packages are currently available and are inexpensive. Popular packages include BILOG-MG,[18] which is capable of fitting a variety of IRT models to MCQ data, as well as WINSTEPS,[19] which is devoted to Rasch modelling more specifically. With regard to IRT models that are better suited to performance assessments, common software packages include FACETS[22] and PARSCALE.[23]

CONCLUSIONS

Both CTT and IRT models have been used extensively over the past decades in a host of medical education assessment programmes, as well as in licensure and certification frameworks. Although IRT is appealing in that the confounding effects of item difficulty and candidate abilities noted in CTT are resolved, it also makes much stronger assumptions and is more complex mathematically. Additionally, minimal sample size requirements for some of the simpler IRT models may not be feasible for some classes.

In instances where the user wishes to simply rank-order students and has little desire to generalise beyond that specific setting (e.g. a clerkship and a specific group of students), or with smaller class sizes, CTT might be sufficient for scoring and other activities. However, class sizes permitting, IRT is useful to answer a host of questions that may arise with medical assessments, both with MCQ examinations and performance assessments.

Both IRT and CTT should be viewed as complementary approaches that can each provide useful information at various phases of activity. For example, preliminary item statistics based on CCT can be useful to identify keying or other procedural errors that may occur in the early phases of processing. Item response theory can subsequently be applied to estimate final item difficulties, candidate ability values, and to complete any other activity that may be necessary to sustain the examination programme.

Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 Downing SM. Item response theory: applications of modern test theory in medical education. *Med Educ* 2003;**37**:739–45.
2 De Champlain A, Melnick DE, Scoles PV *et al.* Assessing medical students' clinical sciences knowledge in France: a collaboration between the National Board of Medical Examiners and a consortium of French medical schools. *Acad Med* 2003;**78**:509–17.
3 Spearman C. Demonstration of formulae for true measurement of correlation. *Am J Psychol* 1907;**18**:161–9.
4 Gulliksen H. *Theory of Mental Tests.* New York, NY: John Wiley 1950.
5 Lord FM, Novick MR. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley 1968.
6 Traub RE, Rowley GL. An NCME instructional module on understanding reliability. *Educ Measure Issue Pract* 1991;**10**:171–9.
7 Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ* 2004;**38**:1006–12.
8 Coombs CH. The concepts of reliability and homogeneity. *Educ Psychol Meas* 1950;**10**:43–56.
9 van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardised patients: state of the art. *Teach Learn Med* 1990;**2**:58–76.
10 Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;**16**:297–334.
11 Subkoviak MJ. A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *J Educ Measure* 1988;**25**:47–55.
12 Harvill LM. An NCME instructional module on standard error of measurement. *Educ Measure Issue Pract* 1991;**10**:181–9.
13 Hambleton RK, Swaminathan H, Rogers J. *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications 1991.
14 Hulin CL, Lissak RI, Drasgow F. Recovery of two and three logistic parameter item characteristic curves: a Monte Carlo study. *Appl Psychol Meas* 1982;**6**:249–60.
15 Wright BD, Stone M. *Best Test Design.* Chicago, IL: MESA Press 1979.
16 Hambleton RK, Jones RW. An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educ Measure Issue Pract* 1993;**12**:253–62.
17 Gessaroli ME & De Champlain A. (2005). Test dimensionality: assessment of. In: Everitt BS, Howell DC, eds. *Encyclopedia of Statistics in Behavioral Science.* Chichester: John Wiley & Sons 2005;2014–21.
18 *BILOG-MG*, Version 3.0. Lincolnwood, IL: Scientific Software International 2003.
19 *WINSTEPS, Version 3.67.0.* Chicago, IL: Winsteps 2008.
20 Wainer H, Dorans NJ, Eignor D, Flaugher R, Green BF, Mislevy RJ, Steinberg L, Thissen D. *Computerized Adaptive Testing: A Primer.* Mahwah, NJ: Lawrence Erlbaum 2000.
21 Ostini R, Nering M. *Polytomous Item Response Theory Models.* Thousand Oaks, CA: Sage Publications 2006.
22 *FACETS, Version 3.65.0.* Chicago, IL: FACETS 2009.
23 Muraki E, Bock D. *PARSCALE 4.* Lincolnwood, IL: Scientific Software International 2008.

Received 21 January 2009; editorial comments to author 4 March 2009; accepted for publication 21 May 2009