Applying the socio-cognitive framework to the BioMedical Admissions Test (BMAT)

Insights from language assessment

Also in this series:

The Impact of High-stakes Examinations on Classroom Teaching: A case study using insights from testing and innovation theory
Dianne Wall

Impact Theory and Practice: Studies of the IELTS test and *Progetto Lingue 2000*
Roger Hawkey

IELTS Washback in Context: Preparation for academic writing in higher education
Anthony Green

Examining Writing: Research and practice in assessing second language writing
Stuart D. Shaw and Cyril J. Weir

Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference, May 2005
Edited by Lynda Taylor and Cyril J. Weir

Examining FCE and CAE: Key issues and recurring themes in developing the First Certificate in English and Certificate in Advanced English exams
Roger Hawkey

Language Testing Matters: Investigating the wider social and educational impact of assessment – Proceedings of the ALTE Cambridge Conference, April 2008
Edited by Lynda Taylor and Cyril J. Weir

Components of L2 Reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners
Toshihiko Shiotsu

Aligning Tests with the CEFR: Reflections on using the Council of Europe's draft Manual
Edited by Waldemar Martyniuk

Examining Reading: Research and practice in assessing second language reading
Hanan Khalifa and Cyril J. Weir

Examining Speaking: Research and practice in assessing second language speaking
Edited by Lynda Taylor

IELTS Collected Papers 2: Research in reading and listening assessment
Edited by Lynda Taylor and Cyril J. Weir

Examining Listening: Research and practice in assessing second language listening
Edited by Ardeshir Geranpayeh and Lynda Taylor

Exploring Language Frameworks: Proceedings of the ALTE Kraków Conference, July 2011
Edited by Evelina D. Galaczi and Cyril J. Weir

Measured Constructs: A history of Cambridge English language examinations 1913–2012
Cyril J. Weir, Ivana Vidaković, Evelina D. Galaczi

Cambridge English Exams – The First Hundred Years: A history of English language assessment from the University of Cambridge 1913–2013
Roger Hawkey and Michael Milanovic

Testing Reading Through Summary: Investigating summary completion tasks for assessing reading comprehension ability
Lynda Taylor

Multilingual Frameworks: The construction and use of multilingual proficiency frameworks
Neil Jones

Validating Second Language Reading Examinations: Establishing the validity of the GEPT through alignment with the Common European Framework of Reference
Rachel Yi-fen Wu

Assessing Language Teachers' Professional Skills and Knowledge
Edited by Rosemary Wilson and Monica Poulter

Second Language Assessment and Mixed Methods Research
Edited by Aleidine J Moeller, John W Creswell and Nick Saville

Language Assessment for Multilingualism: Proceedings of the ALTE Paris Conference, April 2014
Edited by Coreen Docherty and Fiona Barker

Advancing the Field of Language Assessment: Papers from TIRF doctoral dissertation grantees
Edited by MaryAnn Christison and Nick Saville

Applying the socio-cognitive framework to the BioMedical Admissions Test (BMAT)

Insights from language assessment

Edited by

Kevin Y F Cheung
Research and Thought Leadership Group
Cambridge Assessment Admissions Testing

Sarah McElwee
Research and Thought Leadership Group
Cambridge Assessment Admissions Testing

and

Joanne Emery
Consultant
Cambridge Assessment Admissions Testing

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

4843/24, 2nd Floor, Ansari Road, Daryaganj, Delhi – 110002, India

79 Anson Road, #06–04/06, Singapore 079906

www.cambridge.org
Information on this title: www.cambridge.org/9781108439312

First published 2017

20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Printed in

A catalogue record for this publication is available from the British Library

ISBN 978-1-108-43931-2

Contents

5 Making scores meaningful: Evaluation and maintenance of scoring validity in BMAT

Mark Elliott

Cambridge Assessment Admissions Testing

Tom Gallacher

Research and Thought Leadership Group,
Cambridge Assessment Admissions Testing

5.1 Introduction

Previous chapters have considered test taker characteristics, cognitive validity and context validity in relation to BMAT. This chapter concentrates on how aspects of scoring a candidate's responses to BMAT contribute to the test's validity, for both the multiple-choice Sections 1 and 2, and the constructed response marked for Section 3. We outline how 'reliability' is reconceptualised in Weir's (2005) socio-cognitive framework into scoring validity and applied to BMAT. Scoring validity is a wider evaluation of scoring issues than traditional approaches, which separate reliability from validity.

Careful examination of scoring validity sheds light on the operational analyses that are used to monitor BMAT sessions and the steps taken to ensure the integrity of results that are released to universities. Statistical methods are used in the monitoring of BMAT scores, which rely on psychometric models and established forms of evaluation. In particular, data from the sections containing multiple-choice questions (MCQs) is used to calculate statistics that inform test development and evaluation. As psychometric theories underlie the statistics presented and the scoring of MCQ items in BMAT, overviews of Classical Test Theory (CTT) and Rasch analysis are presented to contextualise the discussion of scoring validity in BMAT Sections 1 and 2. Although the issues discussed throughout this chapter necessitate use of statistical terminology, we have kept this to a minimum by describing theories conceptually rather than in technical detail. A more critical examination of these concepts is beyond the scope of this volume and there are many seminal texts that include more nuanced evaluations of these theories (e.g. Andrich 2004, DeVellis 2012, Lord and Novick 1968, Mellenbergh 2011, Rasch 1960/1980).

Instead, the focus of this chapter is on answering questions posed by scoring validity. BMAT's MCQ sections and scoring of BMAT Section 3 are addressed separately, as many of the questions are specific to the scoring of responses in each format. First, we will outline scoring validity as conceptualised by Weir (2005) and further developed by others in language testing (e.g. Geranpayeh 2013, O'Sullivan and Weir 2011), before applying this concept to BMAT.

5.2 Scoring validity and its importance in assessment

Scoring validity 'concerns the extent to which test results are *stable over time*, *consistent in terms of the content sampling,* and *free from bias*' (Weir 2005:23, emphasis in original) – in other words, do the measurement properties and scoring make the candidate's results useful in decision making? Some of these aspects are traditionally (e.g. Lado 1961:31) referred to as 'reliability', and are often discussed alongside a narrow conceptualisation of validity that deals with the aspects covered in other chapters of this volume. Departing from this traditional model, the socio-cognitive framework follows the arguments made by Messick (1989), and conceptualises reliability as part of scoring validity, which is a facet of overall validity. Validity is a property of the inferences drawn from test scores, not a property of the testing instrument in isolation. Therefore, anything that impacts upon the inferences that can be made with a test affects its validity. Scores have limited use when a test is not reliable, while it is equally difficult to draw meaningful inferences from a test that is reliable but does not sample the theorised construct adequately. Within this modern paradigm of validity, a test must demonstrate acceptable levels of 'reliability' as a component of its validity argument. By conceptualising reliability as part of validity rather than a separate characteristic of the assessment, it is not acceptable to argue that a test's shortcomings in either validity or reliability are a consequence of a focus on the other aspect.

Weir's (2005) original conceptualisation of the socio-cognitive framework presented scoring validity as an alternative term for reliability, by extending the traditional approach focused on internal consistency and statistical coefficients, to include marker reliability. Weir included aspects such as rater selection, rater characteristics, the development of criteria/rating scales, the rating process, rating conditions, rater training, rater standardisation and moderation, grading and awarding in the scoring part of his validity framework (see also Shaw and Weir (2007), which presents a fuller treatment of these). In addition to the reliability of the test, scoring validity has thus been expanded to cover other scoring-related test aspects which affect the usefulness of inferences drawn from test scores. For example, Geranpayeh (2013) includes the topics of test difficulty, item discrimination and item bias in discussion of scoring validity in Cambridge English Listening exams. These are key issues to consider when evaluating how MCQs, such as those in BMAT Sections 1 and 2, are scored. The scoring validity component within the socio-cognitive validation framework can be used to pose specific questions for Sections 1 and 2 of BMAT, as follows:

- Are items of appropriate difficulty and do they discriminate between candidates?
- Is there a sufficient level of test reliability?
- Is there any evidence of item bias?
- Do the responses being scored come from the candidate?

For the scoring of written tasks, such as those in BMAT Section 3, a range of other issues must be considered as part of scoring validity, to ensure that examiner marking is free from error (Shaw and Weir 2007):

- Are there clearly defined marking criteria that cover the construct?
- Are markers trained, standardised, checked and moderated?
- Is marking reliable and consistent?

We will now consider each of these questions in turn, in terms of why the question is important for test validity arguments, how they can be answered, and the degree to which BMAT answers that question, before summarising the key issues and concerns. In the following discussion, we consider the scoring validity issues relevant to MCQ sections of BMAT.

5.3 Scoring validity in MCQ sections of BMAT

Psychometric theories such as Classical Test Theory (CTT) and Rasch modelling are commonly used for scoring tests with MCQ items. These theories also provide the basis for statistics that are used to evaluate the performance of tests and items. An outline of these theories is presented in the next two sections to contextualise the statistical coefficients presented later in the chapter, and to provide a summary of how Sections 1 and 2 are scored.

Classical Test Theory

CTT is a psychometric theory widely used across most, if not all, areas of applied scale development, evaluation and assessment research (Devellis 2012). While the technical details of CTT will be familiar to many working in test development and validation, a conceptual summary is useful for those less familiar with the theory. The following overview of CTT is adapted from an unpublished doctoral thesis (Cheung 2014); therefore, portions are similar to the conceptual descriptions provided in that text.

CTT conceptually defines the way that a test response for any item or set of items should be interpreted; this response is known as an observed score, which can be the score for a single item, combined items, or an entire test. Novick (1966:1) specifies CTT as the theory that 'postulates the existence of a true score, that error scores are uncorrelated with each other and with true scores and that observed, true and error scores are linearly related'. This definition specifies that any observed item score is composed of the true score, which represents their ability in the trait being tested, and measurement error, which represents the combination of all non-trait related elements which can affect a candidate's test score (for example fatigue, carelessness and lucky guesses). Mathematically presented this refers to the premise that for each item:

$$X = T + \varepsilon$$

(Where X = the candidate's observed score, T = the candidate's true score and ε = error.)

In its strictest form, CTT assumes that error scores are random across items and not correlated with each other, so the error associated with each individual item has a mean of zero across a large sample of responses (DeVellis 2012). In addition, the errors are not correlated with the observed score or the latent true score. By definition, this latent score is unobservable; therefore, the degree of error associated with item responses is estimated using the concept of parallel tests, by treating each item, or set of items, as a mathematically equivalent measure of the same trait. These concepts are particularly important for calculating estimates of reliability, and they also underpin indicators of difficulty and discrimination (how well the item differentiates between test takers of low and high ability). CTT analysis is conducted immediately after a BMAT session; often on a number of occasions as increasing volumes of test data become available.

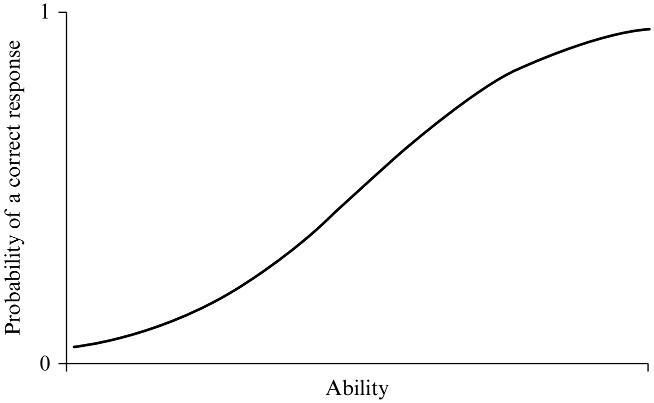
However, CTT has some limitations when applied to test data. One issue arises from the assumption of equal errors across items and respondents. This is unlikely for any test or given testing situation, because the error associated with each item will be different for each respondent (Hambleton, Swaminathan and Rogers 1991). For example, a test administered to a specified sample would include multiple items that vary in difficulty. Taking two items from a test, we could describe one of them as easy to answer correctly and the other one comparatively difficult. Each of these items has associated error that is not fixed across all of the test takers, because the degree of error varies dependent on the ability of the respondent. The easier question will be more useful for discriminating between those of low and mid ability, whereas the more difficult question would more accurately discriminate between those of mid and high ability. Therefore, the amount of error due to guessing is not randomly distributed. Many CTT-based coefficients do not account for these differences and conceptualise error using assumptions that are unlikely to be met precisely in applied testing situations.

Another shortcoming of CTT analyses is that statistics from this framework are descriptive and sample dependent. Therefore they only provide information on how a particular cohort of candidates performed on a particular occasion. The theoretical and practical limitations of CTT have been overcome by complementing CTT analysis with statistics based on a different test theory, known as Item Response Theory (IRT). Cambridge Assessment Admissions Testing uses a specific version of IRT known as Rasch modelling (Rasch, 1960/1980) to score BMAT Sections 1 and 2.

Rasch analysis

The Rasch model (Rasch 1960/1980) is a probabilistic model in which the likelihood of a candidate responding correctly to an item is a function of the difference between the candidate's ability and the *item difficulty*; ability and difficulty are placed on the same scale in units called *logits*. The relationship between ability and the probability of a correct response to a dichotomous item is shown in Figure 5.1, which depicts an *item characteristic curve* (ICC). All Rasch ICCs are parallel, with ICCs for more difficult items located higher (i.e. to the right) along the x axis.

Figure 5.1 Rasch item characteristic curve (ICC)



The Rasch model represents an idealised measurement model which carries the properties of fundamental measurement in the physical sciences, in particular that 'a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or some other occasion' (Rasch 1961:332); in other words, unlike CTT, the Rasch model is sample independent within the parameters of a specified class of items and a specified population – a property known as *specific invariance*. A mathematical feature of the model is that it is possible to condition out person abilities when calculating item difficulty estimates, and vice versa – there is separation of persons and items.

Whereas CTT operates at test level, Rasch analysis focuses at item level, and generates a linear measurement scale. Rasch has advantages over CTT because its results are generalisable beyond the specific test administration, due to the separation of persons and items within the model. Therefore, Rasch-based indicators of item performance, unlike CTT statistics, are not tied to the specific administration of the test from which they were calculated. However, Rasch-based statistics do not supersede CTT ones in operational analysis of BMAT; instead, Rasch- and CTT-based figures are evaluated by data analysts and validation managers as providing complementary information on test performance.

Scoring BMAT MCQ sections

Scores for BMAT Sections 1 and 2 are calculated using the Rasch model. The reported scores are Rasch candidate ability estimates scaled via a linear transformation and reported to one decimal place on a scale that goes from 1.0 to 9.0.

When reporting scores, test providers are concerned with maintaining standards across different versions of the test. For high-stakes testing, using the same test repeatedly is not an option and the test provider must create different forms of the test that are as similar to each other as possible. In language testing, assessments are often benchmarked against external descriptors of language proficiency, such as the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001). In contrast, admissions tests are primarily used to make comparisons within a single cohort of test takers, so the equivalence of BMAT scores between different test years is not a primary concern for selecting institutions. However, it is still necessary to maintain a relatively stable standard from one year to the next for an admissions test. Also, successful applicants may defer their university place for a year and so candidates from two consecutive BMAT test cohorts could be found in a particular *entry* cohort.

Comparability of BMAT scores from consecutive test years is therefore necessary for carrying out predictive validity work within a particular entrant cohort. This comparability is achieved by calibrating scores on the live test cohort using Rasch analysis and then benchmarking them to a subset of the cohort that is regarded as stable in ability. The scaling sets an approximate mean of 5.0 for the subset; scores are capped upwards at 9 and downwards at 1. The scaling ensures that reported test results are comparable within a cohort, with equal intervals in BMAT scale scores representing equal differences in candidate ability. Benchmarking against a stable applicant group allows scores from different sessions to be treated as approximately comparable. However, more precise comparability of test scores is desirable to allow BMAT scores to be used across different test sessions; therefore, Cambridge Assessment Admissions Testing researchers are actively investigating methods of test equating suitable for a high-stakes medical admissions test such as BMAT. This work draws on Cambridge English Language Assessment's expertise in item banking and applying Rasch analysis for test-equating procedures. In addition, relevant experts across Cambridge Assessment's group of exam boards have advised on developments related to scoring; some of the investigations are discussed in the next part.

Developments in scoring BMAT Sections 1 and 2

Recent trials conducted by Cambridge Assessment researchers have explored methods of statistical equating and methods that use expert judgement (e.g. Bramley and Oates 2011) for MCQ sections of BMAT. Rigorous equating methods that allow precise comparability, such as those used with Cambridge English exams, can be achieved by including pretested items when constructing live papers, or by sharing items across sessions. The Cambridge Assessment approach to English language item banking includes anchored pretesting of items, which harnesses the properties of the Rasch model to equate items. Pretesting is a step where candidates that represent the live cohort complete items of known difficulty, referred to as anchor items, along with items whose difficulty is not known; the new items are calibrated through Rasch analysis, using the anchor items to place them on a standardised scale. This produces a bank of pretested items that are available for test construction.

Cambridge English's pretesting model relies on administering items to appropriate cohorts outside of live testing situations. This model has not been applied to BMAT, and while it would provide benefits to test equating across live administrations (contributing to 'parallel forms reliability'), these methods have associated logistical challenges and security issues that need careful consideration before operational deployment. There are some differences between the language testing and admissions testing contexts that influence the viability of pretesting, particularly for a narrowly defined group, such as applicants to study medicine and dentistry. Given these considerations, a range of approaches have been trialled and a number of robust options have been identified. Initially, more than one method will be introduced so that they can be evaluated in parallel. Although this may not be sustainable in the long term, it will provide an evidence base that will allow reliance on a single method of calibrating BMAT scores in the future.

In addition to equating tests and producing scores, psychometric principles are used to analyse test sections that include MCQs. These analyses are employed to evaluate scoring validity of BMAT as part of operational processes, and are outlined throughout the rest of the chapter.

Are items of appropriate difficulty and do they discriminate between candidates?

A test comprised of items which are too difficult or too easy for a candidature will not function well in differentiating between candidates since the scores will be too similar – nearly all correct for an easy test and nearly all incorrect for a hard test. Items need to be of an appropriate difficulty to give a range of scores that reflect the range of abilities within a candidature. If BMAT gave a pass/fail criterion mark based on demonstrating or not demonstrating a fixed level of ability, the most appropriate range of difficulties would be narrowly centred around this level of ability. But since BMAT scores are used by multiple institutions with interests in a range of ability estimates, BMAT questions need to reflect a range of difficulties.

BMAT item writers consider the intended level of test taker abilities when they author items but can naturally never be certain of how difficult their items are until they are taken by real candidates. Basic summary statistics that describe the distribution of scores – the mean and standard deviation – can show whether the overall difficulty of the test was appropriate. If the mean number of items correct is low, it indicates a paper that is too difficult for the candidature and vice versa if the mean is too high. If the standard deviation of scores is too small, then a suitable range of abilities has not been captured.

From a CTT perspective, *item facility* is a statistic that reflects simply the proportion of a candidature that gets an item correct; the higher this value, the easier the item.

$$Item\ facility = \frac{N_{correct}}{N_{total}}$$

(Where $N_{correct}$ is the number of candidates getting the item correct, and N_{total} is the total number of candidates attempting the item.)

Item writers generally intend that 50–60% of their target candidates will get each item correct. As a general rule, items with facility values below 0.1 and above 0.9 are deemed too difficult and too easy respectively to add scoring validity to a test. That is, items cannot discriminate well among candidates if either a very low or very high proportion of candidates get that item correct.

Two items or fewer out of the 62 used in BMAT November 2016, 2015 and 2014 live administrations were outside of the ranges of 0.1 to 0.9, with mean item facilities of 0.47, 0.48 and 0.52 respectively. This indicates how BMAT items are consistently set at appropriate difficulty levels for the cohorts.

From a Rasch perspective, item difficulties need to map reasonably closely to candidate abilities in order to maximise precision of the test for the candidature. By plotting histograms of the distributions of candidate ability estimates and contrasting these with histograms of the distributions of item difficulty estimates, one can visually inspect the appropriateness of a set of items for a set of candidate abilities. Figure 5.2 and Figure 5.3 illustrate the spread of abilities of candidates ('PERSONS': above the line), and the difficulties of the items ('ITEMS': below the line) drawn from a Rasch analysis of BMAT Sections 1 and 2 from November 2016. These confirm that the items were set at appropriate difficulty levels for the cohort. CTT and Rasch analyses of difficulty are routinely used to train item writers in estimating the difficulty of newly authored items.

Figure 5.2 Item difficulty and person ability estimates from BMAT Section 1, November 2016

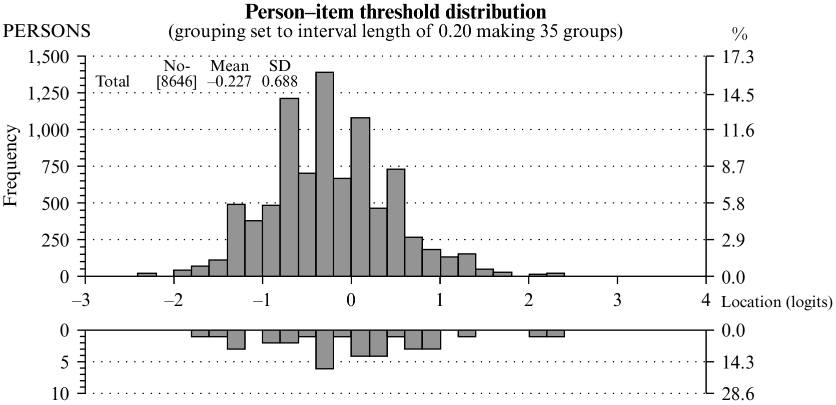
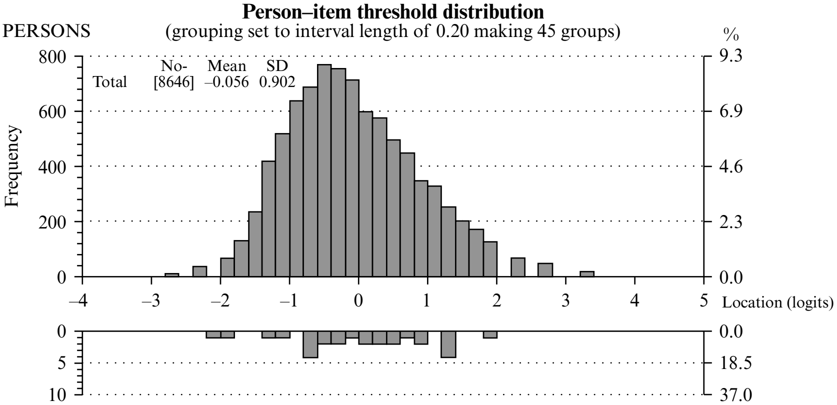


Figure 5.3 Item difficulty and person ability estimates from BMAT Section 2, November 2016



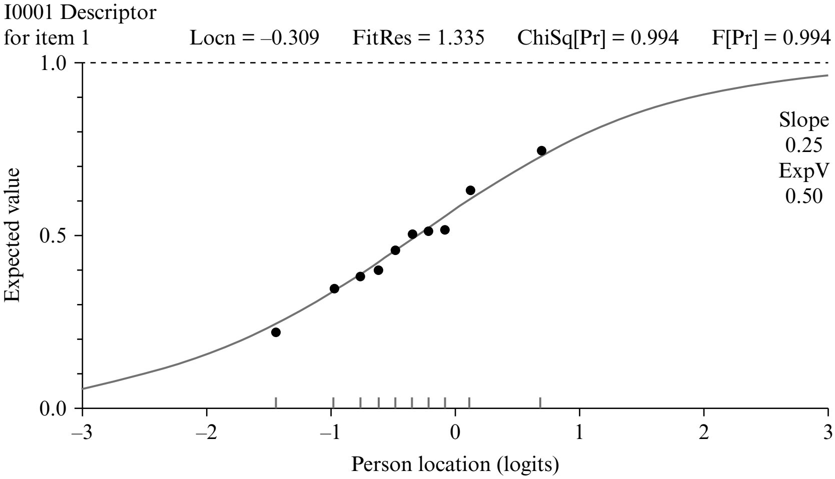
Having covered difficulty, we now turn to *discrimination* of the test items, which refers to how well the items differentiate between test takers with high ability and those with low ability. One simple CTT measure, the *discrimination index*, involves dividing the cohort into three groups (high performers, medium performers and low performers) based on their total score on that paper, then calculating the difference between the facility of each item for the high-scoring group and the low-scoring group. Discrimination index values range from 1 (perfect discrimination) to -1 (perfect negative discrimination). If the discrimination index is too low, it indicates that the item does not discriminate well between candidates among the cohort. Typically, discrimination index values should be 0.3 or higher.

A more sophisticated CTT index, the *point-biserial correlation*, represents the correlation between a candidate's overall test score, and the likelihood of choosing one of the responses. This approach conceptualises the overall test score as an indicator of the candidate's ability. Candidates with higher overall scores will be expected to choose the correct response more often and the incorrect distractor responses less often, compared with those test takers who have lower abilities. The most important point-biserial correlation is therefore the correct response, for which point-biserial values less than 0.25 generally indicate poorly discriminating items. Poor discrimination can be a result of extreme facility values – items which are answered correctly or incorrectly by nearly all candidates naturally cannot discriminate effectively – but can also occur if an item does not assess the same ability as the other items in the set.

As well as a test of the quality of the items in a test, the point-biserial index also forms a key part of the quality assurance process during BMAT marking as a means of checking the correctness of the answer key. All items in BMAT Sections 1 and 2 are multiple-choice format, and any distractor options that have a higher point-biserial correlation than the correct option are flagged for scrutiny, since this may indicate that the given key is incorrect (the flagged option is chosen more often as ability increases, and more so than the given key).

Discrimination is treated somewhat differently in a Rasch analysis; one starts with the premise that all items discriminate, and tests this assumption against the data. Therefore, discrimination is conceptualised as an issue of how well the observed data fits the Rasch model. Figure 5.4 demonstrates the principles of an item where the Rasch model fits the data well. Along the x axis is the 'person' ability estimate in logits, and along the y axis is the expected probability that a candidate will get an item correct. As the ability estimate increases (to the right), the likelihood of that candidate answering the item correctly increases (up). The line represents the theoretical ICC and each black dot represents a group of candidates with similar abilities. For an item for which the Rasch curve is found to fit, the dots for each group should be on or close to the line of the theoretical curve.

Figure 5.4 Item characteristic curve for an item with good fit



For a BMAT session, ICCs are plotted for a visual inspection of each item's fit. Easier items will have lower difficulty estimates, and the ICC will appear to the left, to indicate that groups of candidates with lower abilities are expected to achieve the correct response. Harder items will have higher difficulty estimates, shifting the curve to the right, indicating that only the higher ability candidates are expected to get the item correct.

The degree of divergence between the dots and the line is reported as a mean square statistic for each item, which has a corresponding chi-square and probability value indicating whether the divergence is significantly different from zero. These form the basis for model fit statistics that are also used to establish each item's degree of misfit, and the nature of any issues that are detected. Different types of misfit can provide diagnostic information for BMAT assessment managers and item writers to review as part of question paper production procedures.

Is there a sufficient level of test reliability?

The concept of reliability in a test relates to the stability of test scores. Theoretically, a test with perfect reliability would result in a candidate achieving an identical score every time they took the test, because the score would represent the true ability of the candidate, free from any sources of error. Of course, this hypothetical ideal is not achievable in practice, rendering the premise untestable because this conceptualisation relies on the candidates' abilities remaining precisely stable across administrations of the test. Therefore, reliability of a test is estimated from the consistency of scores for the same candidate, either between different administrations of a test (test–retest reliability) or within a test (internal consistency).

Test–retest reliability

Test–retest reliability involves the administration of the same form of a test to a sample of candidates on two separate occasions and correlating the scores obtained from the two administrations. This method of estimating reliability has several drawbacks (Ebel and Frisbie 1991:81–82), in particular, the restriction to one form does not reflect variation in items across different forms of the test, and candidates are seeing the same items for a second time, so the responses during the second administration will inevitably be influenced by the first due to memory, meaning that the two administrations are not independent. If the time interval between the two administrations is too great, a process of additional learning or attrition is likely to change the candidates' ability, reducing the validity of the comparison. To date, no test–retest reliability studies have been carried out on BMAT.

Parallel form/equivalent form reliability

Parallel form reliability involves administering two different forms of a test in immediate succession to a sample of candidates and correlating the scores. In order to be considered *parallel forms*, not only must the tests be constructed to the same specifications but the scores on the two tests must have the same mean and standard deviation; otherwise, the two forms are referred to as *equivalent forms*.

To date, no studies on parallel forms or equivalent forms have been carried out on BMAT; this is a possible area for future research, although it can be considered to overlap to a large extent with *internal consistency*, as described in the next section. Data does exist, however, on candidates who have taken BMAT in successive years, which can be considered an example of equivalent form reliability. The usefulness of this data, however, is limited under the administration of BMAT, since candidates only have one opportunity per year to achieve scores. Over the course of a year a number of things might change for each candidate to affect their ability, such as the candidate's level of knowledge or motivation being different across application cycles. As a result, we would not expect a candidate's performance to be the same a year later – the candidate should not be expected to have the same level of ability, meaning that few, if any, meaningful inferences can be drawn from the data.

Internal consistency

The internal consistency of a single administration of a fixed-format multi-item test is usually measured by *Cronbach's alpha* (Cronbach 1951), which focuses on the homogeneity of the responses to items within a test administration. Conceptually, it is the mean of all possible split-half correlations, which are the correlations of candidate scores on two halves of the items in a test – this can be viewed as a version of equivalent form reliability where the two equivalent forms are constructed from the two halves of a single test. Mathematically, Cronbach's alpha represents a conceptualisation of reliability as the proportion of variance in observed scores which is accounted for by variance in true scores, with the remaining variance accounted for by error:

$$\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2}$$

(Where ρ_{XX} = reliability, $\sigma^2{}_T$ = variance of true scores, $\sigma^2{}_X$ = variance of observed scores.)

This figure cannot be calculated directly in practice, however, since true scores cannot be known. As an estimate, the proportion of variance of the total scores comprised of the covariance of items, corrected for bias in the estimator for variance (Bol'shev 2001), is used; the standard formula for calculating Cronbach's alpha is:

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{1=1}^{N} \sigma_i^2}{\sigma_X^2}\right)$$

(Where N = number of items in the test, $\sigma^2{}_i$ = variance of scores on item i (for dichotomous items, this is simply $p(1-p)$, where p = item facility), $\sigma^2{}_X$ = variance of scores on the whole test.)

Cronbach's alpha statistics tell us about the scores on an administration of a test, but they are sample dependent; in particular, they are affected by the spread of abilities in the candidature, as well as the number of items, and are sensitive to violations of unidimensionality (Andrich 2009b:3). Assuming that the amount of error in the test remains constant, an increase in the range of ability of the candidature will result in an increase in the variance of the observed scores and a higher alpha. As a result, a higher or lower Cronbach's alpha may be a function of the candidature rather than the items in the test, meaning that there are limits on the interpretation of alpha. Nonetheless, it remains a useful measure to compare two administrations of a test with a stable candidature across time. A related CTT statistic which has less sample dependence than Cronbach's alpha is the *Standard Error of Measurement* (SEM), which represents a standard deviation of the error present in the test. As such, the SEM can be used to construct confidence intervals around test scores. The SEM of a set of items can be calculated by the following formula:

$$SEM = \sigma\sqrt{1-r}$$

(Where σ = standard deviation of raw scores, r = Cronbach's alpha.)

Table 5.1 shows the Cronbach's alpha and SEM figures for BMAT for the period 2012–16, based on analyses of the cohort applying to the University of Cambridge, which represents a consistent, relatively stable cohort and thus renders year-on-year comparisons more valid.

Table 5.1 Cronbach's alphas and SEMs for BMAT 2012–16

Year	Cronbach's alpha		SEM	
	Section 1	Section 2	Section 1	Section 2
2012	0.55	0.63	2.57	2.28
2013	0.61	0.57	2.61	2.34
2014	0.71	0.79	2.53	2.12
2015	0.70	0.80	2.58	2.26
2016	0.72	0.79	2.63	2.25

As can be seen in Table 5.1, the majority of internal consistency estimates have been above 0.70 for BMAT Sections 1 and 2, particularly in recent years. This is in the lower range of coefficients considered acceptable; however, the use of internal consistency estimates as evidence of test quality is problematic because a number of factors impact on these values. In fact, Cronbach himself expressed dissatisfaction with over-zealous application of his alpha in scale evaluation (Cronbach and Shavelson 2004). Other researchers have expressed similar concerns that alpha is not always the best indicator of reliability, particularly when multidimensionality is observed in test responses (Sjitsma 2009, Zinbarg, Revelle, Yovel and Li 2005). Given that BMAT Section 1 is designed to assess three specified thinking skills, and Section 2 includes knowledge from four subject disciplines, some degree of multidimensionality in these sections is inevitable. However, the broad domain coverage of these sections is an important feature of the test that contributes to the cognitive validity of BMAT. Although estimates of internal consistency could be improved by making BMAT sections more unidimensional, this would be detrimental to the quality of inferences that could be made based on the test (Zumbo and Rupp 2004).

Another way of improving internal consistency would be to increase the number of responses marked in the test sections by increasing the number of items. This would either increase the testing time, or reduce the time available for each item, which would have knock-on effects for the context validity of BMAT (see Chapter 4 for a discussion of speededness within BMAT), or indeed the practicality of administering BMAT. Due to the limitations of Cronbach's alpha, SEM values should be inspected when considering internal consistency. Because score standard deviation is included in the calculation of SEMs, they are less susceptible to sample dependence; therefore, SEMs can be regarded as more useful reflections of test reliability (Tighe, McManus, Dewhurst, Chis and Mucklow 2010). The SEM values for BMAT are in line with other MCQ tests of similar length, although the Section 1 values suggest there is some room for these to be improved on the basis of 35 items. The Cronbach's alpha values reported in Table 5.1 dip below 0.60 in two cases and progressively improve for more recent years, whereas the SEM values are relatively consistent throughout, which is encouraging.

One issue to note is that some biomedical schools using BMAT aggregate Section 1 and Section 2 scores to provide a composite score based on all 62 items across these sections. When based on multiple sets of items, rather than a single set, internal consistency can be higher purely as a function of the greater number of items included. For example, composite alphas calculated across all 62 items for the most recent test sessions were 0.83 (2016) and 0.81 (for both 2015 and 2014). This is within the acceptable range for similar tests, indicating that combined aggregates of the sections have good internal consistency. However, the values should be interpreted with caution because Cronbach's alpha is calculated on the basis that individual items contribute to an overall score equally. This is not the case when Rasch scaled scores for the two sections are aggregated together; therefore, these values are merely a rough indicator of the improved internal consistency when scores are calculated using a larger number of items.

In the language testing context, Cambridge English has historically adopted a construct validity approach to developing examinations, unlike other exam boards more heavily influenced by the US psychometric tradition, such as the Educational Testing Service (ETS). Weir (2005:31) describes the focus on internal consistency forms of reliability as 'a fetish with internal consistency among the professional testing fraternity' and points out that very high internal consistency may not be an appropriate aim for tests that seek to evaluate complex and multi-faceted constructs. Over a decade ago, Weir also observed that ETS was acknowledging context validity more readily in the Test of English as a Foreign Language (TOEFL) than they had in the past. Interestingly, a similar change is currently happening in admissions testing contexts, as the SAT has recently been redesigned to have a greater focus on content and learning than has previously been the case (College Board 2015).

Given the narrow ability range of the candidature, the number of test items and, importantly, the broad domain coverage of Sections 1 and 2, the internal consistency of the test sections is acceptable, and respectable when aggregating Sections 1 and 2 together. The reliability of the test sections is routinely monitored in analysis alongside other features, such as item bias, which is evaluated using the procedures described in the following portion of this chapter.

Is there any evidence of item bias?

A test may be considered biased when it produces *systematic* differential performance among test takers of comparable ability on the construct, but who differ on a non-test-related dimension (e.g. in terms of age, gender, race and ethnicity, or physical disability). Bias is a clear threat to test score validity because it prevents the conclusion that ability estimates reflect only the relevant and desired constructs. The BMAT test construction process mitigates bias in test items through the avoidance of culturally bound or sensitive topics and words; however, empirical analysis can still flag individual items that have produced inconsistent performance across different subgroups of test takers. Differential Item Functioning (DIF) analyses (Holland and Thayer 1988, Holland and Wainer (Eds) 1993) can be used to monitor evidence of bias by gender and by school type.

DIF analysis formalises the question of bias by asking whether candidates in a 'focal group', who are indicated to be different on a non-test-related dimension such as gender or school type, have the same probability of getting an item correct in comparison with candidates in a 'reference group' while controlling for ability. The total test score is treated as an indicator of candidate ability for these procedures, which have been employed by Cambridge Assessment researchers with BMAT. An example of this work is presented as a key study below.

Key study – Investigating item bias in BMAT using DIF analysis (Emery and Khalid 2013a)

In the study described here the BMAT performance of different candidate groups (male versus female, independent versus state school) was investigated at the individual item level using the Mantel-Haenszel (MH) procedure (Mantel and Haenszel 1959). The aim was to look for any evidence of DIF in BMAT items by gender and by school sector over multiple years of the test.

Research question

Is there any evidence of DIF by gender or by school sector in BMAT Sections 1 and 2?

Data collection

Candidate-level information (gender and centre number) was matched to BMAT item-level data for test years 2010, 2011 and 2012 (whole cohorts). School type was matched to candidates' centre number (as outlined in Chapter 2). All UK school types other than 'independent' and 'other' were classed as belonging to the state sector. The data of candidates from non-UK schools and those from UK school type 'other' were omitted from school sector analyses. All candidates were included in the gender analyses (see Table 5.2).

Analysis

DIF analyses were carried out using the MH statistical procedure (Holland and Thayer 1988), which uses ability matching by treating the observed total

Table 5.2 Sample sizes for DIF analysis

BMAT year	Gender analysis (N)	School sector analysis (N)
2010	6,225	4,633
2011	6,230	4,681
2012	7,044	4,556

test score as a criterion. In this case, BMAT Section 1 score was used as the criterion for Section 1 items and BMAT Section 2 score was used as the criterion for Section 2 items in each test year. The MH procedure compares the odds of getting the item correct for the reference and focal groups at a given level of ability.

For gender analyses the reference group was defined as male and the focal group as female. For school sector analyses the reference group was defined as independent sector and the focal group as state sector. The following guidelines (Zwick and Ercikan 1989) were used to evaluate the DIF effect size:

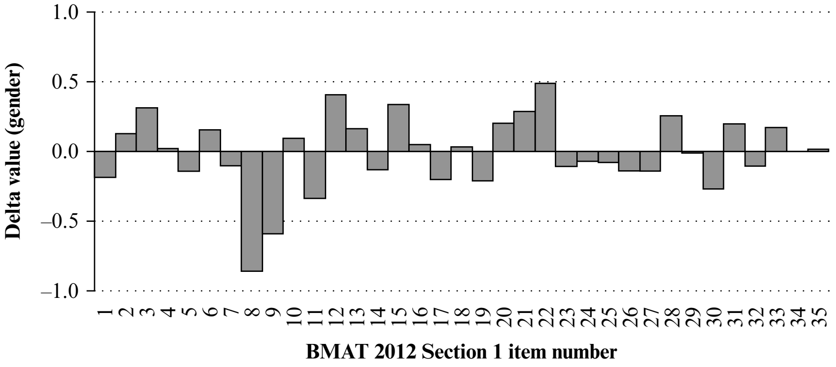
- type A items – negligible DIF/functioning properly: items with |delta value| < 1
- type B items – moderate DIF: items with |delta value| between 1 and 1.5
- type C items – large DIF: items with |delta value| > 1.5.

A negative delta value indicates that the item favours the reference group over the focal group and a positive delta value indicates that the item favours the focal group over the reference group. Delta value thresholds of 1.0 and 1.5 (or -1.0 and -1.5) are equivalent to odds ratios greater than 1.53 (or less than 0.65) and greater than 1.89 (or less than 0.53), respectively. Type A items are considered to function properly but type B and C items require necessary revision and action (Holland and Thayer 1988, Zwick and Ercikan 1989).

Results and discussion

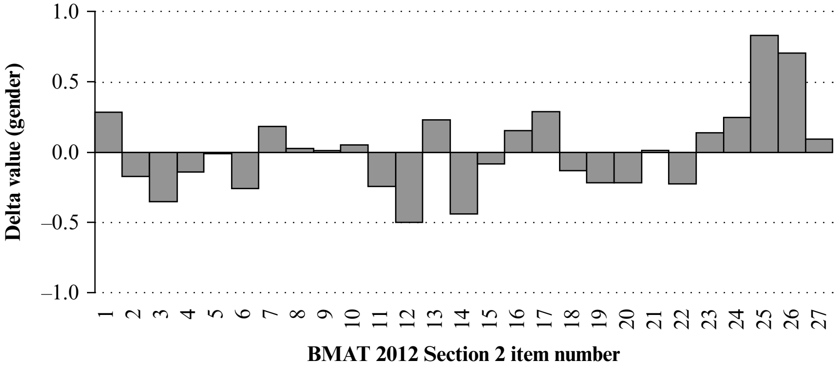
Figure 5.5 and Figure 5.6 display example DIF statistics by gender for BMAT Sections 1 and 2 in the 2012 test. The 2012 figures are presented here as these contained the larger delta values found in the study. No delta value in any of these test years was greater than 1 (or less than -1), indicating no instance of DIF by gender. Additionally, there was no gender pattern evident in the delta values by either item position in the paper or item type (e.g. biology, physics). Four items (out of 186) across all three test years had delta values approaching 1 or -1. These were further scrutinised and were found to belong to a mixture of item subtypes.

Figure 5.5 DIF statistics by gender for BMAT 2012 (Section 1) items*



**A positive delta value indicates that the item favours the focal group (females); a negative delta value indicates that the item favours the reference group (males); delta values 1 to 1.5 = moderate DIF; delta values >1.5 = large DIF.*

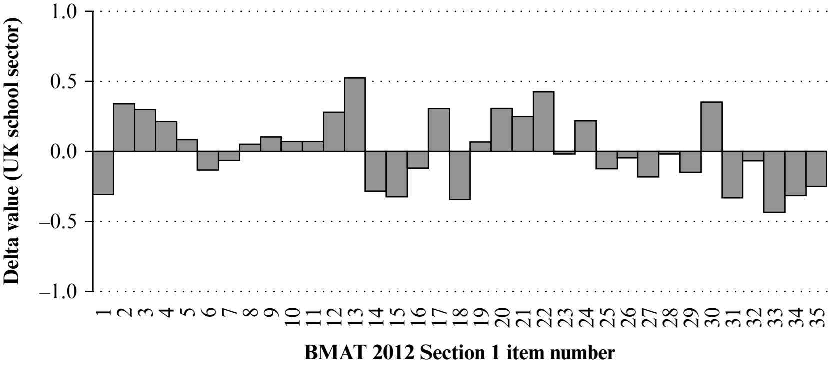
Figure 5.6 DIF statistics by gender for BMAT 2012 (Section 2) items*



**A positive delta value indicates that the item favours the focal group (females); a negative delta value indicates that the item favours the reference group (males); delta values 1 to 1.5 = moderate DIF; delta values >1.5 = large DIF.*

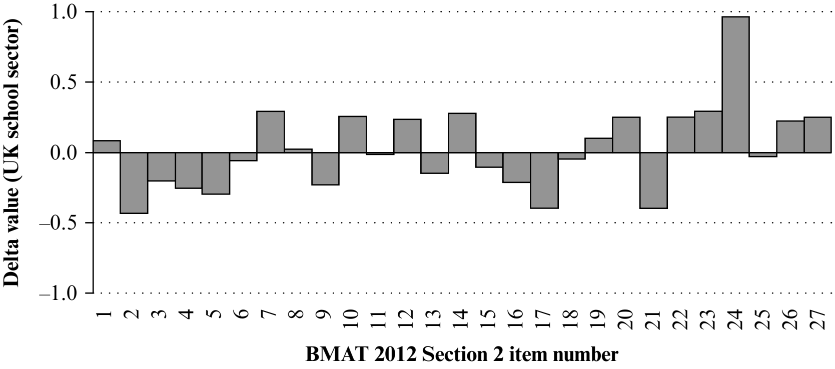
Figure 5.7 and Figure 5.8 display DIF statistics by school sector for BMAT Sections 1 and 2 in the 2012 test. As for gender, no DIF was evident by school sector in any of these test years (no delta value was greater than 1 or less than -1). Again, no pattern in delta values was evident by either item position or item type and very few items in the school sector analyses yielded delta values in excess of 0.5. A single item across all three test years had a delta value close to 1. This was a mathematics question in BMAT 2012 (item 24 in Figure 5.8), which trended towards favouring state school candidates.

Figure 5.7 DIF statistics by UK school sector for BMAT 2012 (Section 1) items*



**A positive delta value indicates that the item favours the focal group (state); a negative delta value indicates that the item favours the reference group (independent); delta values 1 to 1.5 = moderate DIF; delta values >1.5 = large DIF.*

Figure 5.8 DIF statistics by UK school sector for BMAT 2012 (Section 2) items*



**A positive delta value indicates that the item favours the focal group (state); a negative delta value indicates that the item favours the reference group (independent); delta values 1 to 1.5 = moderate DIF; delta values >1.5 = large DIF.*

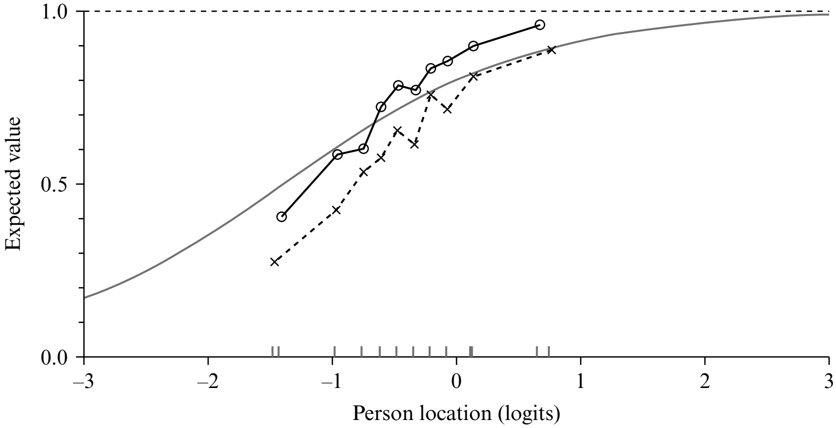
In these three years of BMAT data there was no evidence of DIF by gender or by UK school sector using the MH procedure. Emery and Khalid's (2013a) study informed current Cambridge Assessment practices, which include DIF analyses of BMAT as part of routine operational procedures.

However, the contemporary approach uses more advanced Rasch-based procedures, which are outlined briefly in the next section.

Current Cambridge Assessment practice: Item bias

In Rasch terms, DIF analysis tests whether an item has a significantly different difficulty estimate when treated as two separate items, one for each group (Andrich 2009a, Linacre 2016, Thissen, Steinberg and Wainer 1993). That is to say that for any given ability, one group would have a lower proportion of correct responses than the other. This would manifest in ICCs with two separate lines when plotting observed scores for response categories, as illustrated in Figure 5.9. When the ICCs for two groups are plotted parallel to each other side by side, the group represented by the curve on the right found the item more difficult than the group represented by the curve on the left. This indicates that a person of the same overall ability had a different probability of answering the item correctly, associated with their membership of a particular group.

Figure 5.9 Example ICC for an item exhibiting DIF



Examining each item in a test sequentially can identify items whose content might be related to the non-test-related dimension, potentially indicating bias in the items identified. Criteria for flagging items displaying different degrees of DIF were developed by Zwick, Thayer and Lewis (1999) based on the MH method, with Rasch-based equivalences outlined by Linacre (2016:422). These criteria are predicated on the detected DIF being both statistically significant and substantive in terms of its magnitude; under Linacre's criteria, an item is deemed to display moderate to large DIF (i.e. potentially be a cause for concern) if the magnitude of the DIF is at least 0.64

logits and a significance test on the magnitude of the DIF being 0.43 logits or greater is significant at $p = 0.05$. Here, 0.43 and 0.64 logits equate to MH measures of 1 and 2 δ units respectively, based on an equivalence of 1 logit = 2.35 δ units. A significant DIF result does not automatically indicate that an item is unfair, but it does represent grounds for qualitative investigation by subject experts.

The operational DIF analyses of BMAT November sessions covering the period 2013–16 investigating DIF by gender (male versus female) and school type (state versus independent) were reviewed. They produced three items out of a total of 186 displaying moderate DIF. Items with negligible DIF were balanced between those slightly favouring males and those slightly favouring females. Similarly, items were equally balanced between those slightly favouring independent school candidates and those slightly favouring state school candidates (all with negligible DIF). Assessment managers checked the content of all items flagged at these levels to confirm that there were no task features that give an advantage to one group over another. These analyses are conducted immediately following a test session as part of BMAT's quality assurance procedures.

Are the candidates' responses their own?

Measures such as reliability, error and dimensionality are only meaningful if candidates' test scores reflect their ability in the trait under investigation, and are not the result of issues such as pre-exposure to test items or collusion – in other words, that the integrity of the test is preserved. Ensuring that items are not pre-exposed is a question of test material security, and relates mainly to administrative factors. These issues are discussed in Chapter 4 as contextual features of BMAT's administration. Detecting collusion and copying, on the other hand, is a task that statistical analysis contributes to, alongside monitoring of test centres.

Once BMAT MCQs have been scored, a statistical analysis of candidates' response strings is conducted to identify cases with unusually strong patterns of common wrong answers (instances where pairs of candidates have both chosen the same incorrect option for an item) using Angoff's (1974) A index, following standardised procedures (Bell 2015, Geranpayeh 2014). Angoff's A index compares the proportion of common wrong answers between each pair of candidates relative to their overall score, against the pattern observed across the whole candidature; high indices indicate a greater degree of similarity than that observed between other pairs, which may indicate that collusion or copying has occurred. Often, detection of unusual patterns coincides with reports of unusual activity from exam invigilators. The results of statistical analysis are not treated as definitive proof of malpractice, but rather as an indicator; any cases which are flagged from this check are referred to a malpractice panel for scrutiny, which consists of a senior manager from each of three divisions[1] with some responsibility for BMAT. They are joined by one representative from the biomedical or dentistry departments using BMAT and another Cambridge Assessment colleague who is not involved in the development of admissions tests. The panel reviews the results of statistical analysis on a case-by-case basis, alongside other information, such as seating plans and reports from the exams officers at test centres. In some cases, statements are requested from the candidates and further investigations are conducted to gather information, before final decisions on whether to withhold results are made.

5.4 Scoring validity in BMAT Section 3

Are there clearly defined marking criteria that cover the construct?

BMAT Section 3 responses are marked by two markers against the criteria presented in Table 5.3. Scripts are identified only by BMAT number and candidate initials, so markers are blind to demographic information about the candidate. Each marker gives two scores to each response: one for quality of content (on a scale of 0–5) and one for quality of written English (on the scale A, C, E). Candidates have access to the Writing Task marking criteria on the BMAT website.

In arriving at their scores, markers are instructed to consider whether the candidate has:

- Addressed the question in the way demanded?
- Organised their thoughts clearly?
- Used their general knowledge and opinions appropriately?
- Expressed themselves clearly using concise, compelling and correct English?

The marking criteria against which markers judge the essays are presented in Table 5.3.

Prior to 2010, Writing Task responses were given a single, holistic score that reflected the overall quality of the response. The marking criteria incorporated both the content and the quality of written English descriptors above. The mark scheme was altered in 2010, at the request of stakeholder universities, over concerns that some markers may give more weight than others to quality of written English. The new mark scheme was trialled before its first use (Shannon and Scorey 2010) to ensure that examiners were able to apply it as well as the previous scheme, using a re-marking of a sample of scripts from the live 2009 test.

[1] These are Assessment; Validation and Data Services; and Stakeholder Relations.

Table 5.3 BMAT Section 3 marking criteria

Quality of content

Score	Criteria
5	An excellent answer with no significant weaknesses. ALL aspects of the question are addressed, making excellent use of the material and generating an excellent counter proposition or argument. The argument is cogent. Ideas are expressed in a clear and logical way, considering a breadth of relevant points and leading to a compelling synthesis or conclusion.
4	A good answer with few weaknesses. ALL aspects of the question are addressed, making good use of the material and generating a good counter proposition or argument. The argument is rational. Ideas are expressed and arranged in a coherent way, with a balanced consideration of the proposition and counter proposition.
3	A reasonably well-argued answer that addresses ALL aspects of the question, making reasonable use of the material provided and generating a reasonable counter proposition or argument. The argument is relatively rational. There may be some weakness in the force of the argument or the coherence of the ideas, or some aspect of the argument may have been overlooked.
2	An answer that addresses most of the components of the question and is arranged in a reasonably logical way. There may be significant elements of confusion in the argument. The candidate may misconstrue certain important aspects of the main proposition or its implication or may provide an unconvincing or weak counter proposition.
1	An answer that has some bearing on the question but which does not address the question in the way demanded, is incoherent or unfocused.
0	An answer judged to be irrelevant, trivial, unintelligible or missing will be given a score of 0.

Quality of written English

Band	Criteria
A	Good use of English. Fluent. Good sentence structure. Good use of vocabulary. Sound use of grammar. Good spelling and punctuation. Few slips or errors.
C	Reasonably clear use of English. There may be some weakness in the effectiveness of the English. Reasonably fluent/not difficult to read. Simple/unambiguous sentence structure. Fair range and appropriate use of vocabulary. Acceptable grammar. Reasonable spelling and punctuation. Some slips/errors.
E	Rather weak use of English. Hesitant fluency/not easy to follow at times. Some flawed sentence structure/paragraphing. Limited range of vocabulary. Faulty grammar. Regular spelling/punctuation errors. Regular and frequent slips or errors.
X	A response that is judged to be below the level of an E will receive an X.

The double marks for each response are combined as follows to give the final mark for each candidate. If the two marks for quality of content are the same or no more than one mark apart on the scale, the candidate is awarded the average of the two marks. If the two marks for quality of written English are the same or no more than one mark apart on the scale, the scores are combined like this: AA = A, AC = B, CC = C, CE = D and EE = E. For example, a response given a 4C by one examiner and 4A by the other will get a final score of 4B. A response given 3C by one examiner and 2C by the other will receive a mark of 2.5C.

If there is a larger discrepancy in the marks for either scale then the response is blind-marked for a third time by an experienced marker. The third marker considers only the scale on which the initial discrepancy occurred. If the third mark is the same as, or adjacent to, either of the first two marks then the mean of those two marks is reported. Where the third mark is equally spaced between the first two marks then the mean of all three marks (i.e. the third mark) is reported. All responses awarded a 0 for quality of content or an X for quality of written English are reviewed by an assessment manager to establish whether it deserved such a low mark.

In addition to the Writing Task scores, a scanned image of the response is supplied to the applicant's institution(s). This image provides institutions with a basis for further qualitative assessment of the applicant's writing skills as well as a potential tool for promoting discussion at interview.

Are markers trained, standardised, checked and moderated?

All examiners are recruited based on qualifications (with a minimum of a degree or equivalent), as well as skills and experience set out by recruitment guides appropriate for the level of examiner seniority. More experienced senior examiners are responsible for groups of markers, so the most experienced markers lead those who are less experienced. The more experienced examiners, such as Principal Examiners, are also involved in training more junior ones alongside assessment managers. This approach is informed by studies showing that differences between scores awarded by experienced and inexperienced markers can be reduced through training and standardisation (e.g. Weigle 1999).

New potential markers are required to engage in a ranking exercise. After reading the marking criteria applicants are given eight scripts covering the full range of marks. They are asked to rank these in order of quality of content and to assign a mark on a 3-point scale for quality of English. The aim is to verify that the marker is able to judge the relative quality of responses, although they are not expected to accurately apply the marking criteria at this stage. Assessment managers and Principal Examiners recruit the marking team based on whether potential markers' rankings deviate significantly from what is expected, but also based on interviews.

Marker training and standardisation

Recruited markers attend a compulsory training and standardisation day each year, where markers read through the information available to candidates, the marking criteria and an extract from the official BMAT preparation book, *Preparing for the BMAT* (Shannon (Ed) 2010). Mark schemes have been shown to have a standardising effect on the scores awarded by examiners (Furneaux and Rignall 2000), so the mark scheme for Section 3 is emphasised throughout these sessions, and markers are required to familiarise themselves with it as part of their training. Markers then sit the Writing Task paper under normal test conditions. The markers' own papers are shared out for marking followed by a brief discussion to reflect on the experience of writing their response, and to raise any queries about the application of the marking criteria. These group exercises are used because writing examiners learn mark schemes from their peers and contemporaries (Weigle 1994).

Following the BMAT test session, experienced assessment managers review a large number of scripts to identify examples of candidates' writing that match particular points in the mark scale, both in terms of quality of content and quality of English. The aim is to put together two sets of scripts that contain a representative spread of quality, and for which there is close agreement in terms of the marks that each script should be awarded.

As iterative standardisation has been shown to improve marker reliability (Furneaux and Rignall 2000), two rounds of standardisation exercises are used to prepare markers before marking of live papers begins. First, all markers mark the same four scripts covering a range a quality. Marks are collated and, for each script, a discussion is held of the marks awarded and how these relate to the mark scheme. Where necessary, guidance is given on the interpretation and application of the marking criteria. Another standardisation exercise then requires all markers to mark the same eight scripts. These scripts are chosen to represent a range of quality but include some that may prove difficult to mark (for example, responses where it might be questioned whether the candidate has addressed all of the demands of the question). Again, marks are collated and a discussion held on how they relate to the mark scheme, with any further necessary guidance given.

Marker checking and monitoring

During the live marking period, markers work in teams located together on the same table, led by a Team Leader. These teams are mixed regularly to include experienced markers alongside those who are less experienced. First, second and third marking is all done 'blind' where markers are unaware of the marks given by other markers. Markers take a script, write their marker number on the script, and record marks awarded on a separate sheet. Another marker will take another script and do the same, while yet others might provide second or third marks, and progress the scripts along trays located on the table, indicating the stage of marking. Principal Examiners and Team Leaders submit an evaluation of examiners, indicating the degree of their satisfaction.

Marker behaviour is also monitored statistically, using a 'means model' technique devised by Bell, Bramley, Claessen and Raikes (2007). The mark distribution (mean and standard deviation) of each individual marker is compared to that of the whole group to flag suspected severity, leniency and variability. This method of live marker monitoring assumes that scripts are assigned to markers at random and that, once a minimum number have been marked, we would expect the average mark they award and its standard deviation to be approximately the same as those of all other markers. Markers whose mean score is much lower or higher than others might be exhibiting harshness or leniency, whilst their standard deviation reveals erraticism or, conversely, their failure to use the whole mark range. An assessment manager from Cambridge Assessment oversees this process and answers any queries about the interpretation and application of the mark scheme.

The assessment manager also gives feedback to markers who are flagged, and to markers who are frequently involved in double-marking disagreements. Where there are concerns about a marker's performance, the marker can be dismissed and their scripts re-marked. A feedback session is held at the end of each marking day to address any remaining issues, but markers are not provided information on their marking distributions. There is some evidence that continuous standardisation and feedback can result in a see-saw effect as markers attempt to over adjust (Shaw 2002); therefore, the monitoring for BMAT Section 3 is designed so that assessment managers can control the flow of feedback provided.

From 2017 onwards, Principal Examiners also deal with results enquiries received after results have been released. These were historically marked by the chief examiner for BMAT Section 3, who is always a member of staff from one of the medical or dental schools using BMAT. In the event of an appeal following completion of the results enquiry process, the chief examiner for BMAT Section 3 evaluates all of the previously awarded grades alongside the submission. This arrangement remains unchanged.

Is marking reliable and consistent?

With regard to the consistency of scoring for BMAT Section 3, the *Standards* (2014) make it clear that, since the responses are scored with subjective judgement, evidence should be provided on both *inter-rater consistency* in scoring and (if applicable) *within-examiner consistency* over repeated measurements. In order to unpack that idea we can talk about the need for human markers to be consistent in two different ways: each marker needs to be internally consistent, i.e. given a particular quality of performance, a marker needs to award the same mark whenever this quality appears (*intra-rater reliability*); there also needs to be consistency of marking between markers, i.e. one marker will award the same mark to a constructed response as another marker when confronted with a performance of the same quality (*inter-rater reliability*).

Following the marking period, the level of agreement between the first and second marks is calculated using cross-tabulations, frequencies of mark differences and the percentage of responses that required a third mark. An example cross-tabulation is provided in Table 5.4 and Table 5.5, for BMAT 2015.

Table 5.4 BMAT November 2015 Section 3 marker agreement: Quality of content*

Quality of content		Marker 2						Total
		0	1	2	3	4	5	
Marker 1	0	0	0	0	0	0	0	0
	1	0	64	65	3	0	0	132
	2	0	46	1,087	801	29	1	1,964
	3	0	2	756	3,221	840	9	4,828
	4	0	0	10	858	875	88	1,831
	5	0	0	0	7	95	76	178
Total		0	112	1,918	4,890	1,839	174	8,933

** 59.6% exact agreement (dark shading), 99.3% within one score band (light and dark shading).*

Table 5.5 BMAT November 2015 Section 3 marker agreement: Quality of English*

Quality of English		Marker 2				Total
		X	E	C	A	
Marker 1	X	0	0	0	0	0
	E	0	30	65	4	99
	C	0	46	647	818	1,511
	A	0	7	702	6,613	7,322
Total		0	83	1,414	7,435	8,932

** 81.6% exact agreement (dark shading), 99.9% within one score band (light and dark shading).*

Multi-faceted Rasch analysis (Linacre 2014) is an extension of the Rasch model that can, in addition to person ability and item difficulty, account for additional facets that are seen in scoring that involves judgement. In the case of BMAT Section 3, an analysis was conducted using 2016 data, covering the following facets: 1) candidate ability; 2) marker leniency/severity; and 3) marking criteria (quality of English and quality of content). Among other things, the analysis provides estimates of markers' relative leniency/severity to identify problematic markers, as well as measures of marker consistency.

The multi-faceted Rasch model for BMAT Section 3 exhibited good model fit, explaining over 87% of the variance. Test takers were reliably separated into three strata. As candidates are marked on a 5-point scale in this section of BMAT, one might expect that candidates would be divided into more strata. That they are only separated into just three strata can be explained by the strong BMAT cohort not generally getting scores at the bottom of the scale, as can be seen in Table 5.5.

Quality of markers' marking was generally good. In terms of severity and leniency, markers showed acceptable levels of variation, with the most severe and lenient markers deviating from the mean by less than a quarter of a score point. Only a small number of markers (six of 88) had an infit mean square above 1.5, indicating that the vast majority of markers were marking consistently (Bond and Fox 2001). For the small number of cases where markers are shown to not mark consistently, the findings of these analyses are provided to assessment managers, who provide further training and additional supervision during future marking sessions. If a marker is erratic in their awarding of marks, the assessment manager can decide to dismiss them or exclude them from future marking exercises.

The appeals process allows candidates a re-mark of any section of their BMAT paper. For example, in the November 2015 administration of BMAT, 448 sections were re-marked, with slightly more requests for Section 3 than Sections 1 or 2. After the re-marks for Sections 1 and 2 there were no amendments to the scores from those issued, and nine amendments to Section 3 marks, one of which led to a change greater than one point: the re-mark resulted in a decrease of 1.5 marks in the candidate's quality of content score. The outcomes of the appeals process help to confirm the scoring is valid, but the process also reassures candidates that the mark they receive is the most accurate reflection of their performance that is possible.

5.5 Chapter summary

In this chapter we have answered questions that derive from Weir's (2005) socio-cognitive framework in terms of the scoring validity of BMAT Sections 1, 2 and 3. This was achieved by describing the core analyses that form part of the BMAT yearly cycle, and by presenting results from additional investigations into BMAT's scoring validity. We have also shown how statistics from CTT and Rasch analysis can inform our understanding of test scores, and the importance of understanding the strengths and weaknesses of these models.

Although this chapter emphasises the technical aspects of test validity, it has also presented a critical evaluation of the role that statistical analysis plays in monitoring of a test such as BMAT. Whilst the questions addressed by scoring validity are important, evaluation of the answers in a meaningful way is not possible unless other aspects of validity are considered. Therefore, the methods presented in this chapter and the coefficients they produce should be considered as tools that enable validity to be investigated fully. Many of the statistics outlined favour longer tests with simple items that assess a restricted and simple construct. A test with good fit, discrimination and reliability could be constructed by identifying a simple discriminating item and posing it to test takers repeatedly in slightly different, but many, forms. This approach would certainly be easier and less resource intensive than the production processes outlined in the previous chapter on context validity. However, it would be difficult to favourably judge the construct coverage and theoretical rationale for an admissions test produced in this way. Furthermore, an approach focused on improving psychometric coefficients cannot produce an assessment that satisfactorily evaluates the ability to develop complex arguments.

Cambridge Assessment Admissions Testing's approach, much like that of Cambridge English Language Assessment, is to consider scoring validity alongside other aspects of test validity; this guards against reliably assessing a construct that is not valid or fit for purpose. In particular, test developers have a responsibility to consider scoring validity alongside the rationale for assessing targeted constructs (cognitive validity) and the social impact of bias in assessment (consequential validity), which cannot be achieved without an understanding of the test taker characteristics in a candidature. Weir's (2005) observation that scoring validity is essential, but not sufficient, for presenting an overall validity argument certainly applies to the admissions testing context. In the following chapter, another topic crucial to overall validity of an admissions test is explored – that of criterion-related validity.

Chapter 5 main points

- Coefficients calculated on Sections 1 and 2 of BMAT indicate acceptable levels of psychometric quality.
- Section 3 is assessed by two markers, and Rasch analysis is used to provide evidence of scoring validity for Section 3.
- Statistics can also support detection of malpractice within test cohorts.
- Statistics based on CTT and Rasch analysis can be useful tools for investigating scoring validity of an assessment.
- Commonly reported statistics have specific limitations and weaknesses that should be considered when interpreting them in relation to tests.

References

Admissions Testing Service (2016a) *BMAT Section 1 Question Guide*, available online: www.admissionstestingservice.org/images/324081-bmat-section-1-question-guide.pdf

Admissions Testing Service (2016b) *Biomedical Admissions Test (BMAT) Test Specification*, available online: www.admissionstestingservice.org/images/47829-bmat-test-specification.pdf

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1966) *Standards for Educational and Psychological Testing*, Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1985) *Standards for Educational and Psychological Testing*, Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (2014) *Standards for Educational and Psychological Testing*, Washington, DC: American Educational Research Association.

Anastasi, A and Urbina, S (1997) *Psychological Testing*, New York: Macmillan.

Andrich, D A (2004) Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care* 42 (1), 1–15.

Andrich, D A (2009a) *Interpreting RUMM2030 Part I: Dichotomous Data*, Perth: RUMM Laboratory.

Andrich, D A (2009b) *Interpreting RUMM2030 Part VI: Quantifying Response Dependence in RUMM*, Perth: RUMM Laboratory.

Angoff, W H (1974) The development of statistical indices for detecting cheaters, *Journal of the American Statistical Association* 69 (345), 44–49.

Arthur, N and Everaert, P (2012) Gender and performance in accounting examinations: Exploring the impact of examination format, *Accounting Education: An International Journal* 21 (5), 471–487.

Association of American Medical Colleges (2014) *Core Competencies for Entering Medical Students*, available online: www.staging.aamc.org/initiatives/admissionsinitiative/competencies/

Association of American Medical Colleges (2016) *Using MCAT® Data in 2017 Medical Student Selection*, available online: www.aamc.org/download/462316/data/2017mcatguide.pdf

Atkinson, R C and Geiser, S (2009) Reflections on a century of college admissions tests, *Educational Researcher* 38 (9), 665–676.

Bachman, L (1990) *Fundamental Considerations in Language Testing,* Oxford: Oxford University Press.

Bachman, L and Palmer, A (1996) *Language Testing in Practice*, Oxford: Oxford University Press.

Baldiga, K (2014) Gender differences in willingness to guess, *Management Science* 60, 434–448.

Ball, L J (2014) Eye-tracking and reasoning: What your eyes tell about your inferences, in Neys, W D and Osman, M (Eds) *New Approaches in Reasoning Research*, Hove: Psychology Press, 51–69.

Ball L J and Stupple, E J N (2016) Dual-reasoning processes and the resolution of uncertainty: The case of belief bias, in Macchi, L, Bagassi, M and Viale, R (Eds) *Cognitive Unconscious and Human Rationality*, Cambridge: MIT Press, 143–166.

Barrett, G V, Phillips, J S and Alexander, R A (1981) Concurrent and predictive validity designs: A critical reanalysis, *Journal of Applied Psychology* 66, 1–6.

Bax, S (2013) The cognitive processing of candidates during reading tests: Evidence from eye-tracking, *Language Testing* 30 (4), 441–465.

Bell, C (2015) A modern perspective on statistical malpractice detection, *Research Notes* 59*,* 31–35.

Bell, J F (2007) Difficulties in evaluating the predictive validity of selection tests, *Research Matters* 3, 5–9.

Bell, J F, Bramley, T, Claessen, M J A and Raikes, N (2007) Quality control of examination marking, *Research Matters* 4*,*18–21.

Bell, J F, Judge, S, Parks, G, Cross, B, Laycock, J F, Yates, D and May, S (2005) The case against the BMAT: Not withering but withered? available online: www.bmj.com/rapid-response/2011/10/31/case-against-bmat-not-withering-withered

Ben-Shakhar, G and Sinai, Y (1991) Gender differences in multiple-choice tests: The role of differential guessing tendencies, *Journal of Educational Measurement* 28, 23–35.

Best, R, Walsh, J L, Harris, B H J and Wilson, D (2016) UK Medical Education Database: An issue of assumed consent [Letter to the editor], *Clinical Medicine* 16 (6), 605.

Black, B (2008) *Critical Thinking – a definition and taxonomy for Cambridge Assessment: Supporting validity arguments about Critical Thinking assessments administered by Cambridge Assessment*, Paper presented at 34th International Association of Educational Assessment Annual Conference, Cambridge, 9 September 2008, available online: www.cambridgeassessmentjobs.org/Images/126340-critical-thinking-a-definition-and-taxonomy.pdf

Black, B (2012) An overview of a programme of research to support the assessment of critical thinking, *Thinking Skills and Creativity* 7 (2), 122–133.

Blanden, J and Gregg, P (2004) Family income and educational attainment: A review of approaches and evidence for Britain, *Oxford Review of Economic Policy* 20 (2), 245–263.

Bol'shev, L N (2001) Statistical estimator, in Hazewinkel, M (Ed) *Encyclopedia of Mathematics*, New York: Springer, available online: www.encyclopediaofmath.org/index.php/Statistical_estimator

Bond, T G and Fox, C M (2001) *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, Mahwah: Lawrence Erlbaum.

Borsboom, D, Mellenbergh, G J and van Heerden, J (2004) The concept of validity, *Psychological Review* 111 (4), 1,061–1,071.

Bramley, T and Oates, T (2011) Rank ordering and paired comparisons – the way Cambridge Assessment is using them in operational and experimental work, *Research Matters* 11*,* 32–35.

Bramley, T, Vidal Rodeiro, C L and Vitello, S (2015) *Gender differences in GCSE*, Cambridge: Cambridge Assessment internal report.

Bridges, G (2010) Demonstrating cognitive validity of IELTS Academic Writing Task 1, *Research Notes* 42, 24–33.

Briggs, D C (2001) The effect of admissions test preparation: Evidence from NELS:88, *Chance* 14 (1), 10–18.

Briggs, D C (2004) Evaluating SAT coaching: Gains, effects and self-selection, in Zwick, R (Ed) *Rethinking the SAT: The Future of Standardized Testing in University Admissions*, London: Routledge, 217–234.

British Medical Association (2009) *Equality and Diversity in UK Medical Schools*, London: British Medical Association.

Buck, G, Kostin, I and Morgan, R (2002) *Examining the Relationship of Content to Gender-based Performance Differences in Advanced Placement Exams*, College Board Research Report 2002-12, ETS RR-02-25, Princeton: Educational Testing Service.

Butler, H A (2012) Halpern critical thinking assessment predicts real-world outcomes of critical thinking, *Applied Cognitive Psychology* 25 (5), 721–729.

Butterworth, J and Thwaites, G (2010) *Preparing for the BMAT: The Official Guide to the BioMedical Admissions Test*, Oxford: Heinemann.

Cambridge Assessment (2009) *The Cambridge Approach: Principles for Designing, Administering and Evaluating Assessment*, Cambridge: Cambridge Assessment, available online: www.cambridgeassessment.org.uk/Images/cambridge-approach-to-assessment.pdf

Cambridge English (2014) *Instructions for Secure Administration of Admissions Tests*, Cambridge: UCLES.

Cambridge English (2016) *Principles of Good Practice: Research and Innovation in Language Learning and Assessment*, Cambridge: UCLES, available online: www.cambridgeenglish.org/images/22695-principles-of-good-practice.pdf

Cambridge International Examinations (2016) *Cambridge International AS and A Level Thinking Skills*, available online: www.cie.org.uk/images/329504-2019-syllabus.pdf

Chapman, J (2005) *The Development of the Assessment of Thinking Skills*, Cambridge: UCLES.

Cheung, K Y F (2014) *Understanding the authorial writer: A mixed methods approach to the psychology of authorial identity in relation to plagiarism*, unpublished doctoral thesis, University of Derby.

Cizek, G J (1999) *Cheating on Tests: How to Do It, Detect It, and Prevent It*, London: Lawrence Erlbaum.

Cizek, G J (2012) Defining and distinguishing validity: Interpretations of score meaning and justifications of test use, *Psychological Methods* 17 (1), 31–43.

Cleary, T A (1968) Test bias: Prediction of grades of Negro and white students in integrated colleges, *Journal of Educational Measurement* 5, 115–124.

Cleland, J A, French, F H and Johnston, P W (2011) A mixed methods study identifying and exploring medical students' views of the UKCAT, *Medical Teacher* 33 (3), 244–249.

Cleland, J, Dowell, J S, McLachlan, J C, Nicholson, S and Patterson, F (2012) *Identifying best practice in the selection of medical students (literature review and interview survey)*, available online: www.gmc-uk.org/Identifying_best_practice_in_the_selection_of_medical_students.pdf_51119804.pdf

Coates, H (2008) Establishing the criterion validity of the Graduate Medical School Admissions Test (GAMSAT), *Medical Education* 42, 999–1,006.

College Board (2015) *Test Specifications for the Redesigned SAT*, New York: College Board.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

Cronbach, L J (1951) Coefficient alpha and the internal structure of tests, *Psychometrika* 16 (3), 297–334.

Cronbach, L J (1998) *Essentials of Psychological Testing*, New York: Harper and Row.

Cronbach, L J and Shavelson, R J (2004) My current thoughts on coefficient alpha and successor procedures, *Educational and Psychological Measurement* 64 (3), 391–418.

Department for Education (2014) *Do academies make use of their autonomy?*, available online: www.gov.uk/government/uploads/system/uploads/attachment_data/file/401455/RR366_-_research_report_academy_autonomy.pdf

Department of Labor, Employment and Training Administration (1999) *Testing and Assessment: An Employer's Guide to Good Practices,* Washington, DC: Department of Labor, Employment and Training Administration.

DeVellis, R F (2012) *Scale Development: Theory and Applications* (3rd edition), London: Sage Publications.

Devine, A and Gallacher, T (2017) *The predictive validity of the BioMedical Admissions Test (BMAT) for Graduate Entry Medicine at the University of Oxford*, Cambridge: Cambridge Assessment internal report.

Dowell, J S, Norbury, M, Steven, K and Guthrie, B (2015) Widening access to medicine may improve general practitioner recruitment in deprived and rural communities: Survey of GP origins and current place of work, *BMC Medical Education* 15 (1), available online: bmcmededuc.biomedcentral.com/track/pdf/10.1186/s12909-015-0445-8?site=bmcmededuc.biomedcentral.com

Downing, S M (2002) Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine* 77, S103–S104.

Downing, S M (2003) Validity: On the meaningful interpretation of assessment data, *Medical Education* 37, 830–837.

Du Plessis, S and Du Plessis, S (2009) A new and direct test of the 'gender bias' in multiple-choice questions, *Stellenbosch Economic Working Papers* 23/09, available online: ideas.repec.org/p/sza/wpaper/wpapers96.html

Dunbar, K and Fugelsang, J (2005) Scientific thinking and reasoning, in Holyoak, K J and Morrison, R G (Eds) *The Cambridge Handbook of Thinking and Reasoning*, Cambridge: Cambridge University Press, 705–725.

Dweck, C S (2012) *Mindset: Changing the Way You Think to Fulfil Your Potential*, London: Little, Brown Book Group.

Ebel, R L and Frisbie, D A (1991). *Essentials of Educational Measurement* (5th edition), Englewood Cliffs: Prentice-Hall.

Eccles, J S (2011) Gendered educational and occupational choices: Applying the Eccles et al model of achievement-related choices, *International Journal of Behavioral Development* 35, 195–201.

Eccles, J S, Adler, T F, Futterman, R, Goff, S B, Kaczala, C M, Meece, J L and Midgley, C (1983) Expectations, values, and academic behaviors, in Spence, J T (Ed) *Achievement and Achievement Motives: Psychological and Sociological Approaches*, San Francisco: W H Freeman, 75–146.

Elliot, J and Johnson, N (2005*) Item level data: Guidelines for staff*, Cambridge: Cambridge Assessment internal report.

Elliott, M and Wilson, J (2013) Context validity, in Geranpayeh, A and Taylor, L (Eds) *Examining Listening: Research and Practice in Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press, 152–241.

Elston, M A (2009) *Women and medicine: The future. A report prepared on behalf of the Royal College of Physicians*, available online: www.learning.ox.ac.uk/media/global/wwwadminoxacuk/localsites/oxfordlearninginstitute/documents/overview/women_and_medicine.pdf

Emery, J L (2007a) *A report on the predictive validity of the BMAT (2004) for 1st year examination performance on the Veterinary Medicine course at the University of Cambridge*, Cambridge: Cambridge Assessment internal report.

Emery, J L (2007b) *A report on the predictive validity of the BMAT (2005) for 1st year examination performance on the Medicine and Veterinary Medicine course at the University of Cambridge*, Cambridge: Cambridge Assessment internal report.

Emery, J L (2007c) *Analysis of the relationship between BMAT scores, A level points and 1st year examination performance at the Royal Veterinary College (2005 entry)*, Cambridge: Cambridge Assessment internal report.

Emery, J L (2010a) *A Level candidates attaining 3 or more 'A' grades in England 2006-2009*, Cambridge: Cambridge Assessment internal report.

Emery, J L (2010b) *An investigation into candidates' preparation for the BioMedical Admissions Test (2007 session): A replication involving all institutions*, Cambridge: Admissions Testing Service internal report.

Emery, J L (2013a) *Are BMAT time constraints excessive?*, Cambridge: Cambridge English internal report.

Emery, J L (2013b) *BMAT test-taker characteristics and the performance of different groups 2003–2012*, Cambridge: Cambridge English internal report.

Emery, J L and Bell, J F (2009) The predictive validity of the BioMedical Admissions Test for pre-clinical examination performance, *Medical Education* 43 (6), 557–564.

Emery, J L and Bell, J F (2011) Comment on I C McManus, Eamonn Ferguson, Richard Wakeford, David Powis and David James (2011). Predictive validity of the BioMedical Admissions Test (BMAT): An Evaluation and Case Study. Medical Teacher 33 (1): (this issue), *Medical Teacher* 33*,* 58–59.

Emery, J L and Khalid, M N (2013a) *An investigation into BMAT item bias using DIF analysis*, Cambridge: Cambridge English internal report.

Emery, J L and Khalid, M N (2013b) *Construct investigation into BMAT using Structural Equation Modelling*, Cambridge: Cambridge English internal report.

Emery, J L and McElwee, S (2014) *Student perceptions of selection criteria for medical study: Are admissions tests a deterrent to application?*, Cambridge: Cambridge English internal report.

Emery, J L, Bell, J F and Vidal Rodeiro, C L (2011) The BioMedical Admissions Test for medical student selection: Issues of fairness and bias, *Medical Teacher* 33, 62–71.

Evans, J S B T and Ball, L J (2010) Do people reason on the Wason selection task? A new look at the data of Ball et al (2003), *The Quarterly Journal of Experimental Psychology* 63 (3), 434–441.

Evans, J S B T, Barston, J L and Pollard, P (1983) On the conflict between logic and belief in syllogistic reasoning, *Memory and Cognition* 11 (3), 295–306.

Facione, P A (1990) *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction*, California: The California Academic Press.

Facione, P A (2000) The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skill, *Informal Logic* 20 (1), 61–84.

Ferguson, E and Lievens, F (2017) Future directions in personality, occupational and medical selection: myths, misunderstandings, measurement, and suggestions, *Advances in Health Science Education* 22 (2), 387–399.

Field, A (2013) *Discovering Statistics Using IBM SPSS Statistics*, London: Sage.

Field, J (2011) Cognitive validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking,* Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 112–170.

Fisher, A (1990a) *Research into a higher studies test: A summary*, Cambridge: UCLES internal report.

Fisher, A (1990b) *Proposal to develop a higher studies test: A discussion document*, Cambridge: UCLES internal report.

Fisher, A (1992) *Development of the syndicate's higher education aptitude tests*, Cambridge: UCLES internal report.

Fisher, A (2005) '*Thinking skills' and admission to higher education*, Cambridge: UCLES internal report.

Fitzpatrick, A R (1983) The meaning of content validity, *Applied Psychological Measurement* 7 (1), 3–13.

Furneaux, C and Rignall, M (2007) The effect of standardisation-training on rater judgements for the IELTS Writing Module, in Taylor, L and Falvey, P (Eds) *IELTS Collected Papers*, Cambridge: UCLES/Cambridge University Press, Studies in Language Testing Volume 19, 422–445.

Galaczi, E and ffrench, A (2011) Context validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking,* Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 112–170.

Gale, M and Ball, L J (2009) Exploring the determinants of dual goal facilitation in a rule discovery task, *Thinking and Reasoning* 15 (3), 294–315.

Gallacher, T, McElwee, S and Cheung, K Y F (2017) BMAT 2015 test preparation survey report, Cambridge: Cambridge Assessment internal report.

Garner, R (2015) Number of pupils attending independent school in Britain on the rise, figures show, *The Independent*, 30 April 2015, available online: www.independent.co.uk/news/education/education-news/number-of-pupils-attending-independent-schools-in-britain-on-the-rise-figures-show-10215959.html

General Medical Council (2009) *Tomorrow's Doctors: Outcomes and Standards for Undergraduate Medical Education*, available online: www.gmc-uk.org/Tomorrow_s_Doctors_1214.pdf_48905759.pdf

General Medical Council (2011) *The State of Medical Education and Practice in the UK*, London: General Medical Council.

Geranpayeh, A (2013) Detecting plagiarism and cheating, in Kunnan, A J (Ed) *The Companion to Language Assessment*, London: Wiley Blackwell, 980–993.

Geranpayeh, A (2014) Detecting plagiarism and cheating: Approaches and development, in Kunnan, A J (Ed) *The Companion to Language Assessment Volume II*, Chichester: Wiley, 980–993.

Geranpayeh, A and Taylor, L (Eds) (2013) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press.

Gilhooly, K J, Fioratou, E and Henretty, N (2010) Verbalization and problem solving: Insight and spatial factors, *British Journal of Psychology* 101 (1), 81–93.

Gill, T, Vidal Rodeiro, C L and Zanini, N (2015) *Students' choices in Higher Education*, paper presented at the BERA conference, Queen's University Belfast, available online: cambridgeassessment.org.uk/Images/295319-students-choices-in-higher-education.pdf

Goel, V, Navarrete, G, Noveck, I A and Prado, J (2017) Editorial: The reasoning brain: The interplay between cognitive neuroscience and theories of reasoning, *Frontiers in Human Neuroscience* 10, available online: journal.frontiersin.org/article/10.3389/fnhum.2016.00673/full

Goodman, N W and Edwards, M B (2014) *Medical Writing: A Prescription for Clarity*, Cambridge: Cambridge University Press.

Green, A (1992) *A Validation Study of Formal Reasoning Items*, Cambridge: UCLES internal report.

Green, A (2003) *Test impact and English for academic purposes: A comparative study in backwash between IELTS preparation and university professional courses*, Unpublished doctoral dissertation, University of Surrey.

Green, A (2006) Watching for washback: Observing the influence of the International English Language Testing System Academic Writing Test in the classroom, *Language Assessment Quarterly* 3 (4), 333–368.

Green, A (2007) Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses, *Assessment in Education: Principles, Policy and Practice* 1, 75–97.

Green, A (2013) Washback in language assessment, *International Journal of English Studies* 13 (2), 39–51.

Griffin, B and Hu, W (2015) The interaction of socio-economic status and gender in widening participation in medicine, *Medical Education* 49 (1), 103–113.

Halpern, D F (1999) Teaching for critical thinking: Helping college students develop the skills and dispositions of a critical thinker, *New Directions for Teaching and Learning* 80, 69–74.

Hambleton, R K and Traub, R E (1974) The effect of item order on test performance and stress, *The Journal of Experimental Education* 43 (1), 40–46.

Hambleton, R K, Swaminathan, H and Rogers, H (1991) *Fundamentals of Item Response Theory*, Newbury Park: Sage Publications.

Hamilton, J S (1993) *MENO Thinking Skills Service: Development and Rationale*, Cambridge: UCLES internal report.

Hawkey, R (2011) Consequential validity, in Geranpayeh, A and Taylor, L (Eds) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press, 273–302.

Haynes, S N, Richard, D C S and Kubany, E S (1995) Content validity in psychological assessment: A functional approach to concepts and methods, *Psychological Assessment* 7 (3), 238–247.

Hecker, K and Norman, G (2017) Have admissions committees considered all the evidence? *Advances in Health Sciences Education* 22 (2), 573–576.

Hembree, R (1988) Correlates, causes, effects, and treatment of test anxiety, *Review of Educational Research* 58, 47–77.

Hirschfeld, M, Moore, R L and Brown, E (1995) Exploring the gender gap on the GRE subject test in economics, *Journal of Economic Education* 26 (1), 3–15.

Hoare, A and Johnston, R (2011) Widening participation through admissions policy – a British case study of school and university performance, *Higher Education Quarterly* 36, 21–41.

Hojat, M, Erdmann, J B, Veloski, J J, Nasca, T J, Callahan, C A, Julian, E R and Peck, J. (2000) A validity study of the writing sample section of the Medical College Admission Test, *Academic Medicine*, 75, 25S–27S.

Holland, P W and Thayer, D T (1988) Differential item performance and Mantel-Haenszel procedure, in Wainer, H and Braun, I (Eds) *Test Validity*, Hillsdale: Lawrence Erlbaum, 129–145.

Holland, P W and Wainer, H (Eds) (1993) *Differential Item Functioning*, Hillsdale: Lawrence Erlbaum.

Hopkins, K, Stanley, J, Hopkins, B R (1990) *Educational and Psychological Measurement and Evaluation*, Englewood Cliffs: Prentice-Hall.

Hu, L T and Bentler, P (1999) Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modelling* 6, 1–55.

Hughes, A (2003) *Testing for Language Teachers* (2nd edition)*,* Cambridge: Cambridge University Press.

Hyde, J S, Lindberg, S M, Linn, M C, Ellis, A B, and Williams, C C (2008) Gender similarities characterize math performance, *Science* 321, 494–495.

Independent Schools Council (2015) *ISC Census 2015*, available online: www.isc.co.uk/media/2661/isc_census_2015_final.pdf

Independent Schools Council (2016) *ISC Census 2016*, available online: www.isc.co.uk/media/3179/isc_census_2016_final.pdf

James, W and Hawkins, C (2004) Assessing potential: The development of selection procedures for the Oxford medical course, *Oxford Review of Education* 30, 241–255.

Jencks, C and Crouse, J (1982) Aptitude vs. achievement: should we replace the SAT? *The Public Interest* 67, 21–35.

Joint Council for Qualifications (2016a) *Adjustments for candidates with disabilities and learning difficulties: Access arrangements and reasonable adjustments*, available online: www.jcq.org.uk/exams-office/access-arrangements-and-special-consideration

Joint Council for Qualifications (2016b) *General and vocational qualifications: General regulations for approved centres*, available online: www.jcq.org.uk/exams-office/general-regulations

Julian, E R (2005) Validity of the Medical College Admission Test for predicting medical school performance, *Academic Medicine* 80, 910–917.

Kane, M (2013) Validating the interpretations and uses of test scores, *Journal of Educational Measurement* 50, 1–73.

Kaplan, R M and Saccuzzo, D P (2012) *Psychological Testing: Principles, Applications, and Issues*, California: Wadsworth Publishing Company.

Katz, S and Vinker, S (2014) New non-cognitive procedures for medical applicant selection: A qualitative analysis in one school, *BMC Medical Education*, available online: www.ncbi.nlm.nih.gov/pubmed/25376161

Kellogg, J S, Hopko, D R and Ashcraft, M H (1999) The effects of time pressure on arithmetic performance, *Journal of Anxiety Disorders* 13 (6), 591–600.

Kelly, M E, Gallagher, N, Dunne, F and Murphy, A (2014) Views of doctors of varying disciplines on HPAT-Ireland as a selection tool for medicine, *Medical Teacher* 36 (9), 775–782.

Kelly, S and Dennick, R. (2009). Evidence of gender bias in True-False-Abstain medical examinations, *BMC Medical Education,* available online: www.ncbi.nlm.nih.gov/pmc/articles/PMC2702355/

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29. Cambridge: UCLES/Cambridge University Press.

Klahr, D and Dunbar, K (1988) Dual space search during scientific reasoning, *Cognitive Science* 12 (1), 1–48.

Klein, S, Liu, O L, Sconing, J, Bolus, R, Bridgeman, B, Kugelmass, H and Steedle, J (2009) *Test Validity Study (TVS) Report*, Washington, DC: US Department of Education.

Koenig, T W, Parrish, S K, Terregino, C A, Williams, J P, Dunleavy, D M and Volsch, J M (2013) Core personal competencies important to enteringstudents' success in medical school: What are they and how could they be assessed early in the admission process? *Academic Medicine* 88 (5), 603–613.

Kreiter, C D and Axelson, R D (2013) A perspective on medical school admission research and practice over the last 25 years, *Teaching and Learning in Medicine* 25, S50–S56.

Ku, K Y L (2009) Assessing students' critical thinking performance: Urging for measurements using multi-response format, *Thinking Skills and Creativity* 4, 70–76.

Kuncel, N R and Hezlett, S A (2010) Fact and fiction in cognitive ability testing for admissions and hiring decisions, *Current Directions in Psychological Science* (19) 6, 339–345.

Kuncel, N R, Hezlett, S A and Ones, D S (2001) A comprehensive meta-analysis of the predictive validity of the Graduate Records Examinations: Implications for graduate student selection and performance, *Psychological Bulletin* 127, 162–181.

Kusurkar, R A, Ten Cate, T J, van Asperen, M and Croiset, G (2011) Motivation as an independent and a dependent variable in medical education: A review of the literature, *Medical Teacher* 33 (5), 242–262.

Lado, R (1961) *Language Testing: The Construction and Use of Foreign Language Tests. A Teacher's Book*, New York: McGraw Hill.

Landrum, R E and McCarthy, M A (2015) Measuring critical thinking skills, in Jhangiani, R S, Troisi, J D, Fleck, B, Legg, A M and Hussey, H D (Eds) *A Compendium of Scales for Use in the Scholarship of Teaching and Learning*, available online: teachpsych.org/ebooks/compscalessotp

Lawshe, C H (1975) A quantitative approach to content validity, *Personnel Psychology* 28, 563–575.

Leijten, M and Van Waes, L (2013) Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes, *Written Communication* 30 (3), 358–392.

Linacre, J M (2014) *Facets computer program for many-facet Rasch measurement*, version 3.71.4, Beaverton: Winsteps.com.

Linacre, J M (2016) *Winsteps® Rasch Measurement Computer Program User's Guide*, Beaverton: Winsteps.com.

Linn, R L (2009) Considerations for college admissions testing, *Educational Researcher* 38 (9), 677–679.

Liu, O L, Frankel, L and Roohr, K C (2014) Assessing critical thinking in higher education: Current state and directions for next-generation assessment, *ETS Research Report Series* 1, 1–23.

Long, R (2017)GCSE, AS and A Level reform, House of Commons briefing paper Number SN06962, available from: researchbriefings.parliament.uk/ResearchBriefing/Summary/SN06962

Lord, F M and Novick, M R (1968) *Statistical Theories of Mental Test Scores*, Reading: Addison-Wesley.

Lu, Y and Sireci, S G (2007) Validity issues in test speededness, *Educational Measurement: Issues and Practice* 26, 29–37.

Luxia, Q (2007) Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China*, Assessment in Education: Principles, Policy and Practice* 1, 51–74.

Mantel, N and Haenszel, W (1959) Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* 22 (4), 719–748.

Massey, A J (2004) *Medical and veterinary admissions test validation study*, Cambridge: Cambridge Assessment internal report.

Mayer, R E, Larkin, J H and Kadane, J (1984) A cognitive analysis of mathematic problem-solving ability, in Sternberg, R J (Ed) *Advances in the Psychology of Human Intelligence*, Hillsdale: Lawrence Erlbaum, 231–273.

McCarthy, J M and Goffin, R D (2005) Selection test anxiety: Exploring tension and fear of failure across the sexes in simulated selection scenarios, *International Journal of Selection and Assessment* 13 (4), 282–295.

McCurry, D and Chiavaroli, N (2013) Reflections on the role of a writing test for medical school admissions, *Academic Medicine* 88 (5), 568–571.

McDonald, A S (2001) The prevalence and effects of test anxiety in school children, *Educational Psychology* 21 (1) 89–101.

McDonald, R P (1981) The dimensionality of tests and items, *British Journal of Mathematical and Statistical Psychology* 34 (1), 100–117.

McManus, I C, Dewberry, C, Nicholson, S and Dowell, J S (2013) The UKCAT-12 study: Educational attainment, aptitude test performance, demographic and socio-economic contextual factors as predictors of first year outcome in a collaborative study of twelve UK medical schools, *BMC Medicine* 11, available online: bmcmedicine.biomedcentral.com/articles/10.1186/1741-7015-11-244

McManus, I C, Dewberry, C, Nicholson, S, and Dowell, J S, Woolf, K and Potts, H W W (2013) Construct-level predictive validity of educational attainment and intellectual aptitude tests in medical student selection: Meta-regression of six UK longitudinal studies, *BMC Medicine* 11, available online: bmcmedicine.biomedcentral.com/articles/10.1186/1741-7015-11-243

McManus, I C, Powis, D A, Wakeford, R, Ferguson, E, James, D and Richards, P (2005) Intellectual aptitude tests and A Levels for selecting UK school leaver entrants for medical school, *BMJ* 331, 555–559.

Medical Schools Council (2014) *Selecting for Excellence Final Report*, London: Medical Schools Council.

Mellenbergh, G J (2011) *A Conceptual Introduction to Psychometrics. Development, Analysis, and Application of Psychological and Educational Tests,* The Hague: Eleven International Publishing.

Messick, S (1989) Validity, in Linn, R L (Ed) *Educational Measurement* (3rd edition), Washington DC: The American Council on Education and the National Council on Measurement in Education, 13–103.

Messick, S (1995) Validity of psychological assessment: Validation of inferences from person's responses and performance as scientific inquiry into scoring meaning, *American Psychologist* 9, 741–749.

Milburn A (2012) *Fair access to professional careers – A progress report by the Independent Reviewer on Social Mobility and Child Poverty*, London: Cabinet Office.

Morris, B J, Croker, S, Masnick, A M and Zimmerman, C (2012) The emergence of scientific reasoning, in Kloos, H, Morris, B J and Amaral, J L (Eds) *Current Topics in Children's Learning and Cognition*, Rijeka: InTech, 61–82.

Ndaji, F, Little, J and Coe, R (2016) *A comparison of academic achievement in independent and state schools: Report for the Independent Schools Council January 2016*, Durham: Centre for Evaluation and Monitoring, Durham University, available online: www.isc.co.uk/media/3140/16_02_26-cem-durham-university-academic-value-added-research.pdf

Newble, D (2016) Revisiting 'The effect of assessments and examinations on the learning of medical students', *Medical Education* 50 (5), 498–501.

Newble, D I and Jaeger, K (1983) The effect of assessments and examinations on the learning of medical students, *Medical Wducation* 17 (3), 165–171.

Newton, P and Shaw, S D (2014) *Validity in Educational and Psychological Assessment*, London: Sage.

Nicholson, S and Cleland, J (2015) Reframing research on widening participation in medical education: using theory to inform practice, in Cleland, J and Durning, S J (Eds) *Researching Medical Education*, Oxford: Wiley Blackwell, 231–243.

Niessen, A S M and Meijer, R R (2016) Selection of medical students on the basis of non-academic skills: is it worth the trouble? *Clinical Medicine* 16(4), 339–342.

Niessen, A S M, Meijer, R B and Tendeiro, J N (2017) Applying organizational justice theory to admission into higher education: Admission from a student perspective, *International Journal of Selection and Assessment* 25 (1), 72–84.

Norris, S P (1990) Effect of eliciting verbal reports of thinking on critical thinking test performance, *Journal of Educational Measurement* 27 (1), 41–58.

Novick, M R (1966) The axioms and principal results of classical test theory, *Journal of Mathematical Psychology* 3 (1), 1–18.

Nowell, A and Hedges, L V (1998) Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores, *Sex Roles* 39 (1/2), 21–43.

O'Hare, L and McGuiness, C (2009) Measuring critical thinking, intelligence and academic performance in psychology undergraduates, *The Irish Journal of Psychology* 30, 123–131.

O'Hare, L and McGuiness, C (2015) The validity of critical thinking tests for predicting degree performance: A longitudinal study, *International Journal of Educational Research* 72, 162–172.

O'Sullivan, B and Weir, C J (2011) Test development and validation, in O'Sullivan, B (Ed) *Language Testing: Theories and Practices*, Basingstoke: Palgrave Macmillan, 13–32.

Palmer, E J and Devitt, P G (2007) Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? *BMC Medical Education* 7, bmcmededuc.biomedcentral.com/articles/10.1186/1472-6920-7-49

Papp, S and Rixon, S (forthcoming 2017) *Assessing Young Language Learners: The Cambridge English Approach*, Studies in Language Testing volume 47, Cambridge: UCLES/Cambridge University Press.

Patel, V L, Arocha, J F and Zhang, J (2005) Thinking and reasoning in medicine, in Holyoak, K J and Morrison, R G (Eds) *The Cambridge Handbook of Thinking and Reasoning*, Cambridge: Cambridge University Press, 727–750.

Patterson, F, Knight, A, Dowell, J S Nicholson, S., Cousans, and Cleland, J. (2016). How effective are selection methods in medical education? A systematic review, *Medical Education* 50, 36–60.

Paul, R and Elder, L (2007) *Critical Thinking Competency Standards (For Educators)*, Tomales: Foundation for Critical Thinking.

Pearson VUE (2017) *UK Clinical Aptitude Test (UKCAT) Consortium UKCAT Examination Executive Summary Testing Interval: 1 July 2016–4 October 2016*, available online: www.ukcat.ac.uk/media/1057/ukcat-2016-technical-report-exec-summary_v1.pdf

Pelacia, T and Viau, R (2017) Motivation in medical education, *Medical Teacher* 39 (2), 136–140.

Plass, J A and Hill, K T (1986) Children's achievement strategies and test performance: The role of time pressure, evaluation anxiety and sex, *Developmental Psychology* 22 (1), 31–36.

Powis, D A (2015) Selecting medical students: An unresolved challenge, *Medical Teacher* 37 (3), 252–260.

Quality Assurance Agency (2002) *Subject Benchmark Statement: Medicine*, available online: www.qaa.ac.uk/en/Publications/Documents/Subject-benchmark-statement-Medicine.pdf

Quality Assurance Agency (2015) *Subject Benchmark Statement: Biomedical Sciences*, available online: www.qaa.ac.uk/en/Publications/Documents/SBS-Biomedical-sciences-15.pdf

Ramsay, P A (2005) *Admissions tests (Cambridge TSA and BMAT) and disability*, Cambridge: University of Cambridge internal report.

Rasch, G (1960/1980) *Probabilistic Models for Some Intelligence and Attainment Tests*, Chicago: University of Chicago Press.

Rasch, G (1961) On general laws and meaning of measurement in psychology, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (4), Berkeley: University of California Press, 321–333.

Rasch, G (2011) *All statistical models are wrong!*, available online: www.rasch.org/rmt/rmt244d.html

Reibnegger, G, Caluba, H-C, Ithaler, D, Manhal, S, Neges, H M and Smolle, J (2010) Progress of medical students after open admission or admission based on knowledge tests, *Medical Education* 44, 205–214.

Röding, K and Nordenram, G (2005) Students' perceived experience of university admission based on tests and interviews, *European Journal of Dental Education* 9 (4), 171–179.

Rodriguez, M C (2003) Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations, *Journal of Educational Measurement, 40*(2), 163–184.

Ross, J A, Scott, G and Bruce, C D (2012) The gender confidence gap in fractions knowledge: Gender differences in student belief–achievement relationships, *School Science and Mathematics* 112 (5), 278–288.

Sackett, P R and Yang, H (2000) Correction for range restriction: An expanded typology, *Journal of Applied Psychology* 85, 112–118.

Sam, A, Hameed, S, Harris, J, Meeran, K (2016) Validity of very short answer versus single best answer questions for undergraduate assessment, *BMC Medical Education* 16 (1), available online: bmcmededuc.biomedcentral.com/articles/10.1186/s12909-016-0793-z

Saville, N and Hawkey, R (2004) The IELTS impact study: Investigating washback on teaching materials, in Cheng, L, Watanabe, Y and Curtis, A (Eds) *Washback in Language Testing: Research Context and Methods*, London: Lawrence Erlbaum, 73–96.

Saville, N (2003) The process of test development and revision within UCLES EFL, in Weir, C J and Milanovic, M (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press, 57–120.

Saville, N (2012) Applying a model for investigating the impact of language assessment within educational contexts: The Cambridge ESOL approach, *Research Notes* 50, 4–8.

Scardamalia, M and Bereiter, C (1987) Knowledge telling and knowledge transforming in written composition, in Rosenberg, S (Ed) *Advances in Applied Psycholinguistics, Volume 2: Reading , Writing and Language Learning*, Cambridge: Cambridge University Press, 142–175.

Schwartzstein, R, Rosenfeld, G, Hilborn, R, Oyewole, S and Mitchell, K. (2013) Redesigning the MCAT exam: balancing multiple perspectives, *Academic Medicine* 88 (5), 560–567.

Scorey, S. (2009a) *Investigating the predictive validity of the BMAT: An analysis using examination data from the Royal veterinary College BVetMed course for the 2005, 2006 and 2007 BMAT cohorts*, Cambridge: Cambridge Assessment internal report.

Scorey, S (2009b) *Investigating the predictive validity of the BMAT: An analysis using examination data from the University College London course for the 2003 to 2007 BMAT cohorts*, Cambridge: Cambridge Assessment internal report.

Seyan K, Greenhalgh T and Dorling D (2004) The standardised admission ratio for measuring widening participation in medical schools: analysis of UK medical school admissions by ethnicity, socioeconomic status, and sex, *British Medical Journal* 328, 1,545–1,546.

Shannon, M D (2005) *Investigation of possible indictors of excessive time pressure in BMAT*, Cambridge: Cambridge Assessment internal report.

Shannon, M D and Scorey, S (2010) *BMAT Section 3 marking trial March 2010 – Marker reliability analysis*, Cambridge:Cambridge Assessment internal report.

Shannon, M D (2010) (Ed) *Preparing for the BMAT: The Official Guide to the BioMedical Admissions Test*. Oxford: Heinemann.

Sharples, J M, Oxman, A D, Mahtani, K R, Chalmers, I, Oliver, S, Collins, K, Austvoll-Dahlgren, A and Hoffmann, T (2017) Critical thinking in healthcare and education, *BMJ* 357, available online: www.bmj.com/content/357/bmj.j2234.long

Shaw, S D (2002) The effect of standardisation on rater judgement and inter-rater reliability, *Research Notes* 8, 13–17.

Shaw, S D and Weir, C J (2007) *Examining Writing: Research and Practice in Assessing Second Language Writing*, Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.

Shea, J and Fortna, G (2002). Psychometric methods, in Norman, G R, van der Vleuten, C P and Newble, D I (Eds) (2012) *International Handbook of Research in Medical Education (Vol. 7)*, New York: Springer Science and Business Media, 97–126.

Shultz, M M and Zedeck, S (2012) Admission to law school: New measures, *Educational Psychologist* 47 (1), 51–65.

Simon, H A and Newell, A (1971) Human problem solving: The state of the theory in 1970, *American Psychologist* 12 (2), 145–159.

Sireci, S G (1998) The construct of content validity, *Social Indicators Research* 45*,* 83–117.

Sjitsma, K (2009) On the use, misuse, and the very limited usefulness of Cronbach's alpha, *Psychometrika* 74 (1), 107–120.

Soares, J A (2012) The future of college admissions: Discussion, *Educational Psychologist* 47 (1), 66–70.

Stegers-Jager, K M, Steyerberg, E W, Lucieer, S M and Themmen, A P N (2015) *Medical Education* 49 (1), 124–133.

Stemler, S E (2012) What should university admissions tests predict? *Educational Psychologist* 47 (1), 5–17.

Steven, K, Dowell, J S, Jackson, C and Guthrie, B (2016) Fair access to medicine? Retrospective analysis of UK medical schools application data 2009–2012 using three measures of socioeconomic status, *BMC medical education* 16 (1), available online: bmcmededuc.biomedcentral.com/articles/10.1186/s12909-016-0536-1

Stevens L, Kelly M E, Hennessy M, Last J, Dunne F, O'Flynn S (2014) Medical students' views on selection tools for medical school – a mixed methods study, *Irish Medical Journal* 107 (8), 229–231.

Stoet, G and Geary, D C (2013) Sex differences in mathematics and reading achievement are inversely related: within- and across-nation assessment of 10 Years of PISA data, *PLOS ONE*, available online: journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0057988&type=printable

Stupple, E J N, Maratos, F A, Elander, J, Hunt, T E, Cheung, K Y F and Aubeeluck, A V (2017) Development of the Critical Thinking Toolkit (CriTT): A measure of student attitudes and beliefs about critical thinking, *Thinking Skills and Creativity* 23, 91–100.

Tai, R H, Loehr, J F and Brigham, F J (2006) An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments, *International Journal of Research and Method in Education* 29 (2), 185–208.

Taylor, L (Ed) (2011) *Examining Speaking: Research and Practice in Assessing Second Language Speaking,* Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press.

Thissen, D, Steinberg, L and Wainer, H (1993) Detection of differential item functioning using the parameters of item response models, In Holland, P and Wainer, H (Eds) *Differential Item Functioning.* Hillsdale: Lawrence Erlbaum, 67–113.

Thomson, A and Fisher A (1992) *MENO: A validation study of informal reasoning items*, Norwich: University of East Anglia internal report.

Tiffin, P A, McLachlan, J C, Webster, L and Nicholson, S (2014) Comparison of the sensitivity of the UKCAT and A Levels to sociodemographic characteristics: A national study, *BMC Medical Education* 14, available online: bmcmededuc.biomedcentral.com/articles/10.1186/1472-6920-14-7

Tighe, J, McManus, I C, Dewhurst, N G, Chis, L and Mucklow, J (2010) The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP (UK) examinations, *BMC Medical Education* 10, available online: bmcmededuc.biomedcentral.com/articles/10.1186/1472-6920-10-40

Trainor, S (2015) Student data privacy is cloudy today, clearer tomorrow, *The Phi Delta Kappan* 96 (5), 13–18.

Tsai, M-J, Hou, H-T, Lai, M-L, Liu, W-Y and Yang, F-Y (2012) Visual attention for solving multiple-choice science problem: An eye-tracking analysis, *Computers and Education* 58 (1), 375–385.

Universities and Colleges Admissions Service (2016) *Applicant numbers to 'early deadline' university courses increase by 1%, UCAS figures reveal today*, available online: www.ucas.com/corporate/news-and-key-documents/news/applicant-numbers-%E2%80%98early-deadline%E2%80%99-university-courses-increase

Weigle, S C (1994) Effects of training on raters of ESL compositions, *Language Testing* 11 (2), 197–223.

Weigle, S C (1999) Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing* 6 (2), 145–178.

Weigle, S C (2002) *Assessing Writing*, Cambridge: Cambridge University Press.

Weir, C J (2005) *Language Testing and Validation: An Evidence-based Approach*, Basingstoke: Palgrave Macmillan.

Weir, C J and Taylor, L (2011) Conclusions and recommendations, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing Volume 30, Cambridge: UCLES/Cambridge University Press, 293–313.

Wilhelm, O and Oberauer, K (2006) Why are reasoning ability and working memory capacity related to mental speed? An investigation of stimulus–response compatibility in choice reaction time tasks, *European Journal of Cognitive Psychology* 18 (1), 18–50.

Willmott, A (2005) *Thinking Skills and admissions: A report on the validity and reliability of the TSA and MVAT/BMAT assessments*, Cambridge: Cambridge English internal report.

Woolf, K, Potts, H W W, Stott, J, McManus, I C, Williams, A and Scior, K (2015) The best choice? *The Psychologist* 28, 730–735.

Wouters, A, Croiset, G, Galindo-Garre, F and Kusurkar, R A (2016) Motivation of medical students: Selection by motivation or motivation by selection, *BMC Medical Education* 16 (1), available online: www.ncbi.nlm.nih.gov/pubmed/26825381

Wouters, A, Croiset, G, Schripsema, N R, Cohen-Schotanus, J, Spaai, G W G, Hulsman R L and Kusurkar, R A (2017) A multi-site study on medical school selection, performance, motivation and engagement, *Advances in Health Sciences Education* 22 (2), 447–462.

Wright, S (2015) Medical school personal statements: a measure of motivation or proxy for cultural privilege? *Advances in Health Sciences Education* 20, 627–643.

Yeager, D S and Dweck, C S (2012) Mindsets that promote resilience: When students believe that personal characteristics can be developed, *Educational Psychologist, 47*(4), 302–314.

Yu, G, He, L and Isaacs, T (2017). *The Cognitive Processes of taking IELTS Academic Writing Task 1: An Eye-tracking Study*, IELTS Research Reports Online Series, British Council, IDP: IELTS Australia and Cambridge English Language Assessment, available online: www.ielts.org/-/media/research-reports/ielts_online_rr_2017-2.ashx

Zeidner, M (1998) *Test Anxiety: The State of the Art*, New York: Plenum.

Zimmerman, C (2000) The development of scientific reasoning skills, *Developmental Review* 20, 99–149.

Zimmerman, C (2007) The development of scientific thinking skills in elementary and middle school, *Developmental Review* 27, 172–223.

Zinbarg, R E, Revelle, W, Yovel, I and Li, W (2005) Cronbach's α, Revelle's β, and McDonald's ωH: Their relations with each other and two alternative conceptualizations of reliability, *Psychometrika* 70 (1), 123–133.

Zohar, A and Peled, B (2008) The effects of explicit teaching of metastrategic knowledge on low- and high-achieving students, *Learning and Instruction* 18 (4), 337–352.

Zumbo, B D and Rupp, A A (2004) Responsible modelling of measurement data for appropriate inferences: Important advances in reliability and validity theory, in Kaplan, D (Ed) *The SAGE Handbook of Quantitative Methodology for the Social Sciences*, Thousand Oaks: Sage Press, 73–92.

Zwick, R (Ed) (2004) *Rethinking the SAT: The Future of Standardized Testing in University Admissions,* London: Routledge.

Zwick, R and Ercikan, K (1989) Analysis of differential item functioning in the NAEP history assessment, *Journal of Educational Measurement* 26, 55–66.

Zwick, R, Thayer, D T and Lewis, C (1999) An empirical Bayes approach to Mantel-Haenszel DIF analysis, *Journal of Educational Measurement* 36 (1), 1–28.