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

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

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

6 The relationship between test scores and other measures of performance

Molly Fyfe
Research and Thought Leadership Group, Cambridge Assessment Admissions Testing

Amy Devine
Research and Thought Leadership Group, Cambridge Assessment Admissions Testing

Joanne Emery
Consultant, Cambridge Assessment Admissions Testing

6.1 Introduction

This chapter deals with an aspect of validity that is commonly researched in the admissions testing context: criterion-related validity. Criterion-related validation aims to demonstrate that test scores are systematically related to another indicator or outcome ('criterion') that is relevant to the construct measured by the test. It asks how strongly the criterion of interest is related to scores on the test, and is usually investigated using statistics that indicate the strength of relationships, such as correlation and regression.[1]

Weir's (2005) framework conceptualises two types of criterion-related validity: concurrent and predictive. Concurrent validity seeks to establish a relationship between two or more measures of the same ability that are administered at the same time, where the assessment being evaluated is one of the measures. For example, a medical licensing exam given at the end of medical school would be expected to correlate with other robust measures of clinical performance, and relationships between these variables would be taken as evidence of concurrent validity.

[1] Positive correlation coefficients can range from 0 to 1, with higher numbers indicating a stronger relationship between the assessment and the criterion variable. In the context of selection, coefficients above $r = 0\ .35$ are considered very useful, while those below $r = 0.3$ are considered moderately useful, and below $r = 0.1$ as weak (Cleland et al 2012).

Box 6.1 Definition of criterion

The criterion variable is a measure of some attribute or outcome that is operationally distinct from the test. Thus, the test is not a measure of the criterion, but rather is a measure hypothesized as a potential predictor of that targeted criterion.

(*Standards* 2014:17)

Predictive validity seeks to establish a relationship between scores from an assessment and a measure of future performance. The criterion variable used to evaluate the assessment typically becomes available after the test has been administered. Criteria used for predictive validity tend to be measures of different, but theoretically related, constructs to the one represented by a test score. They are often outcomes of interest for reasons other than test validation. In the context of admissions tests, criterion variables used in predictive studies tend to be ones that indicate academic success at university. As admissions tests are used to infer an applicant's potential to be successful in their studies, predictive validity is a particularly important aspect of validity in selection contexts.

For selection tests such as BMAT, fitness for purpose is closely bound with the question of whether test scores differentiate between candidates in a way that relates to their future performance. Selection tests and other selection criteria are forms of 'predictive assessment' (James and Hawkins 2004:241) in that they aim to assess *potential* for a future course of study or job role. According to a previous edition of the *Standards* (1985:11), 'predictive studies are frequently, but not always, preferable to concurrent studies of selection tests for education or employment' whereas 'concurrent evidence is usually preferable for achievement tests, tests used for certification, diagnostic clinical tests, or for tests used as measures of a specified construct'. Establishing good predictive validity is often seen as the holy grail of admissions tests and it is generally accepted that predictive studies should be conducted with selection assessments. Indeed, the published research on validity of admissions tests within medical education has largely focused on investigating the relationship between test scores and measures of future course performance (e.g. Emery and Bell 2009, Emery et al 2011, McManus, Dewberry, Nicholson and Dowell 2013, McManus, Dewberry, Nicholson, Dowell, Woolf and Potts 2013).

The emphasis on predictive validity is greater in admissions testing than it is in language testing. Despite this, Weir's (2005) socio-cognitive framework developed in language testing is useful for framing the criterion-related validity of admissions tests such as BMAT. In the opening chapter of this volume, Saville used the framework to pose the following questions in relation to criterion validity:

- Do test scores relate to other tests or measurements? (concurrent validity)
- Do test scores relate to future outcomes? (predictive validity)

When carrying out studies of criterion-related validity, it is important to acknowledge that longitudinal studies of predictive validity can be more practically difficult to conduct than concurrent studies, and are also more susceptible to influence from confounding variables. Weir (2005:209) points out that 'predictive validity is, however, in general beset with problems because of the variables that may interfere with the comparison over time'. In addition to the challenges of tracking test takers over time, the ways that a test is used can impact greatly on the availability of data, resulting in complications when conducting statistical analysis (see Box 6.2 for examples); these issues will be discussed in depth later in the chapter.

In this chapter we present studies on criterion validity of BMAT conducted by Cambridge Assessment researchers, in light of issues highlighted by Weir's (2005) socio-cognitive framework. In addition, we draw on work conducted in other selection contexts, such as occupational psychology, to outline some of the challenges facing admissions testing researchers. The focus is primarily on predictive validity, although a discussion of relevant concurrent validity considerations is included. To provide a clear picture of the issues, the present chapter begins with a non-technical description of the theoretical and methodological issues related to criterion validity. For each of the topics discussed, Cambridge Assessment's approach to them with BMAT is described. Following this discussion, a key study is described in detail and the findings of other studies are summarised briefly.

6.2 Key issues for investigating criterion-related validity of BMAT

There is a clear rationale for establishing criterion-related validity of educational assessments used in selection contexts (Anastasi and Urbina 1997, *Standards* 2014); however, *how* to conduct criterion-related validation in real-world settings is a somewhat murkier issue. The primary challenge in criterion-related validity is that research must be conducted within the real-world practices of selection. Many best practices within the context of medical education, such as multi-faceted selection procedures[2] or extra support and remediation for specific groups of students, pose methodological challenges for criterion-related validation. There are many

[2] Cambridge Assessment Admissions Testing advocates that BMAT be used alongside other selection criteria, as this is seen as best practice in medical school admissions (Cleland et al 2012); however, a multi-faceted selection process poses methodological challenges for criterion-related validation.

difficulties with investigating criterion-related validity in the context of medical selection; for example, James and Hawkins (2004) list seven specific challenges for evaluating predictive validity in this context (Box 6.2).

Box 6.2 Difficulties associated with investigating predictive validity of medical selection methods (adapted from James and Hawkins 2004:244)

- Selection assessment could produce qualitative or non-normally distributed data.
- The cohort completing the assessment might be too small.
- Assessment scores may not be recorded after decision making.
- The time from administration of the selection assessment to the availability of outcomes is usually very long.
- Outcome variables are only available for successful applicants, who are a subset of the applicants that completed the selection assessment.
- Measures of validity are sensitive to error variation and are dependent on reliability.
- The outcome may neither be reliable nor valid.

With predictive studies, as with most methodologies, a researcher must fully understand the theoretical issues at play, appreciate the situational constraints in studying a real-world phenomenon, and then make and defend a series of expert judgements about how to conduct the desired research. In this part of the chapter, we describe some of the key theoretical and methodological issues in criterion-related validation, and the approach that Cambridge Assessment Admissions Testing takes to researching these areas with BMAT.

Selecting suitable outcome criteria

When considering criterion-related validity, one of the first issues that must be addressed is what criterion a test should be related to. The answer to this question in the context of admissions testing is not clear-cut (Stemler 2012), and researchers must develop a rationale for what they will measure and at what point in time they will measure it. As predictive validity is a key concern of admissions tests, we focus the discussion on selecting outcome criteria for predictive validity studies.

Some researchers approach predictive validity as a purely empirical issue in which any selection variable can be justified if it relates to a desirable outcome. To illustrate this approach, Hopkins, Stanley and Hopkins (1990:82) use an example of selecting employees for sales positions; they argue that: 'If people who indicate that they prefer strawberry to vanilla ice cream become more successful salespeople, then that would be a relevant item for inclusion in the screening test for sales people.'

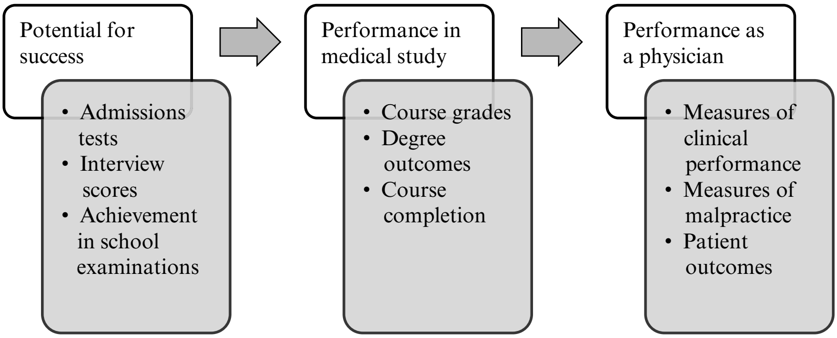
This approach rejects the need for theoretical or logical accounts that attempt to explain the relationships being investigated, particularly when tests are used in selection. Applied to admissions tests, the position advocated by Hopkins et al (1990) typically means that the criterion becomes the focus of validation studies rather than the assessment being validated. For admissions testing researchers adopting this approach, the validity of the outcome measure, normally a grade point average (GPA), is considered self-evident; therefore, any variable that predicts the desirable outcome can be validated as a selection variable.

Cambridge Assessment takes a different position on criterion-related validity in admissions tests. Selecting suitable outcome criteria should be based on the theory behind the test construct, and this is a fundamental consideration in undertaking criterion-related validity studies. Although predictive validity is particularly important in selection contexts, it is still treated as one aspect of overall validity. In addition, we do not treat outcome measures uncritically; instead we recognise that many outcome measures are available in the higher education context and that relationships worthy of investigation should be identified on theoretical grounds. In this regard, we agree with Sireci's (1998:98) observation that: 'Because no one criterion is sufficient for the validation of a test, and because criteria must also be validated, criterion-related studies were only a part of the larger process of construct validation.'

As admissions tests are intended to help select students most likely to be successful in a given endeavour, a theoretical approach to selecting outcome criteria should focus on indicators of success in that endeavour. However, indicators of success can vary on a number of dimensions. Firstly, they can be measured at different points in time, ranging from 'proximal' indicators which are measured fairly close to the assessment, to 'distal' indicators, which are measurements taken far into the future. For example, in professional education, such as to become a doctor or a lawyer, admissions tests could predict success in the first year of study (a proximal outcome) or success as a professional five years after graduation (a more distal outcome). In medical education, there is debate about whether the aim of an admissions test is to select students who have the capabilities to succeed academically on the course or those who will eventually make good doctors. These aims may not be easy to marry and require different criterion measures. Cleland et al (2012:6) point out that attempting to predict who will be a good doctor can be problematic because 'this is a somewhat indeterminate and distal criterion, in the sense that performance as a doctor is not a discrete construct and is temporally distant from selection'.

While there is continuity of development between study and professional practice, the skills needed for success in these endeavours likely differ (Shultz and Zedeck 2012). Figure 6.1 depicts some of the criteria that could be used to assess predictive validity in a medical selection context.

Figure 6.1 Examples of criteria for measuring success in the progression of medical education


Potential for success
• Admissions tests
• Interview scores
• Achievement in school examinations
Performance in medical study
• Course grades
• Degree outcomes
• Course completion
Performance as a physician
• Measures of clinical performance
• Measures of malpractice
• Patient outcomes

While it might be appealing to determine whether an admissions test can predict professional performance, there is great potential for spurious results from this type of analysis due to confounding factors.[3]

> Naturally, the most interesting outcomes to predict would be those that are more distant in time . . . The tension from a psychometric perspective, however is that the greater amount of time that elapses between instruction [or test administration] and assessment, the more mediating variables can creep in that impact on subsequent performance (for good or ill) making it difficult to link outcomes to predictors
> (Stemler 2012:10).

In this regard, Woolf, Potts, Stott, McManus, Williams and Scior (2015) argue that selection for training should also deselect those who are unsuitable for clinical practice, because once applicants are accepted onto a training course nearly all of them qualify to practise. On the other hand, it is expected that the learning and experience provided by the university through teaching and apprenticeship ultimately shapes a trainee's development in the medical

[3] While attempting to establish predictive validity of an admissions test based on distal performance outcomes may be inadvisable, universities may well want to review their selection process as a whole, in light of the proximal and distal performance of their graduates (Stemler 2012).

or dental profession; therefore, screening out individuals before they have benefited from this training needs a strong justification, particularly because non-academic skills have a stronger theoretical relationship with suitability for clinical practice than academic abilities. In relation to this, Niessen and Meijer (2016) have suggested that optimising training of non-academic skills would be preferable to selecting students on the basis of these skills.

In the case of BMAT, the test construct focuses on potential for the course of study (a proximal outcome), rather than predicting who will make a good doctor, although the former (passing the course) is a necessary condition for the latter. There are a number of different ways to look at proximal outcomes related to performance on a medical course. Many would argue that the purpose of admissions tests is to deselect those applicants who are unlikely to succeed and that test scores should not be expected to differentiate between candidates beyond an ‘adequate’ or cut-off level. Others believe that the purpose is selection of the very best candidates in terms of course performance. Whether criterion measures should indicate excellent, adequate or poor performance is therefore a decision to be made in establishing a test’s predictive validity.

BMAT aims to help schools select from the applicant pool those who have a good chance of completing the course of study successfully whilst rejecting those who are least likely to succeed. It is therefore desirable to show that BMAT scores relate to either future course performance itself or to other, known indicators of this. Of course, scores on any selection test only show us what a candidate *could* achieve in the future rather than what they necessarily *will* achieve, which is shaped by many additional factors.

The wider literature on admissions tests shows that first year grade point average (FYGPA) is one of the most common measurements of course performance used for establishing predictive validity. An acknowledged limitation of this approach is that achievement on a course only reveals part of the spectrum of learning and achievement produced through study in higher education (Stemler 2012). Other criteria that might be used to assess predictive validity include rates of attrition and course completion, but these are only useful in contexts where the outcomes occur regularly and are theoretically related to the constructs assessed by a test, such as when students are not progressing due to academic failure.

Cambridge Assessment researchers use measures of academic performance on biomedical courses, including GPA, as a criterion in predictive studies, in line with conventions established by other researchers in educational assessment. We also identify course performance indicators that align theoretically with the cognitive processes and skills assessed by BMAT outlined in Chapter 3. This is done on a case-by-case basis in collaboration with university tutors, in order to acknowledge the complexity of teaching and learning contexts in schools of medicine and dentistry. Commonly this includes grades for particular course components, grades during early years designated as pre-clinical, and course completion when there are concerns about students' abilities to cope with the science-based study required on a course.

Concurrent validity in the admissions testing context

Concurrent validity is a key component of criterion-related validity; however as the purpose of an admissions test is geared towards selecting for good future performance, concurrent validity is considered less often than predictive validity for admissions tests. Additionally, there are a number of challenging issues that arise when considering concurrent validity for a test used in selection for the healthcare professions.

In the language testing context criterion-related validity has a strong focus on concurrent validity. Relating test scores to other well-established measures of language performance helps to establish the validity of an instrument. Concurrent validity is particularly useful when validating a new tool that acts as a more efficient substitute for an established assessment. In this regard Anastasi and Urbina write:

> Because the criterion for concurrent validity is always available at the time of testing, we might ask what function is served by the [new] test in such situations. Basically, such tests provide a simpler, quicker, or less expensive substitute for the criterion data. For example, if the criterion consists of continuous observation of a patient during a two-week hospitalization period, a test that could sort out normal from disturbed or doubtful cases would appreciably reduce the number of persons requiring such extensive observation (Anastasi and Urbina 1997:119).

One of the immediate challenges to conducting concurrent validity studies on BMAT is that there is not an established measure of the same construct available against which to correlate BMAT test scores, nor is there a common framework of standards that define what a medical student should be able to do at entry into medical school[4]. In language testing, there are established frameworks of language proficiency, such as the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001), which support concurrent validation studies by providing a common rubric against which skills measured on different assessments or qualitative measures can be interpreted. A common framework in language testing also

[4] Medical schools frequently use competency frameworks that guide their curriculum and provide detailed statements of what students should be able to do at the end of medical school. However, as these competencies are largely developed through the training provided in medical school, they are not appropriate for establishing a framework of abilities for entering medical students.

supports concurrent validation through 'comparability studies' of different exams that are benchmarked to assess the same levels of the framework.

Medical education currently lacks a common framework that can be used to benchmark potential for medical education. The Association of American Medical Colleges (AAMC, 2014) has recently made a first step in this regard by proposing a series of core competencies for 'entry into medicine' which is being adopted by medical schools in the US and Canada for graduate entry into medical study. The competencies identified by the AAMC link closely to the skills assessed by BMAT sections (see Chapter 3). However, at present, the AAMC's competencies lack structured definitions of levels of ability. In the UK, in which entry into medicine is predominantly at the undergraduate level, there is not yet consensus as to what competencies an entering student should have. Without a common framework, the admissions tests used for medicine and dentistry in the UK (BMAT, United Kingdom Clinical Aptitude Test (UKCAT) and Graduate Medical School Admissions Test (GAMSAT)) conceptualise potential for success at medical school in different ways. From a cognitive validity perspective, the tests are assessing different constructs (although all are described as aspects of potential for biomedical study), and thus comparisons of the scores between these tests would be problematic to interpret.

Of course, potential for success in biomedical study is represented in various ways, not just by admissions test scores. The most commonly used selection criteria are measures of academic achievement at school. Within the broader literature on admissions tests for higher education, secondary school GPA is frequently considered as a criterion for concurrent validity. Admissions test scores and school-based qualifications, such as A Level grades, are typically determined in the late stages of secondary school, so studies examining these variables are often considered as concurrent validity designs (Coates 2008). While one can consider school-leaving qualifications as a concurrent measure for evaluating test validity, a closer look at the timing and constructs in the UK context prompts us to re-examine this position.

Cambridge Assessment Admissions Testing has investigated the relationships between BMAT performance and A Level achievement, which can be conceptualised as a concurrent or predictive criterion. To contextualise Cambridge Assessment's work on BMAT's relationship with A Level grades, we must address the grey area between predictive and concurrent validity.

The *Standards* (2014) pose that 'historically, two designs, often called predictive and concurrent, have been distinguished for evaluation of test-criterion relationships' (2004:17). While researchers tend to agree on distinguishing between these two designs for criterion-related validity, there is debate over where the dividing line is drawn. Traditionally, in occupational psychology settings, the distinction between concurrent and predictive validity designs is based on whether the criterion was measured at the same time as the assessment being validated (Barrett, Phillips and Alexander 1981). In contrast, Anastasi and Urbina state:

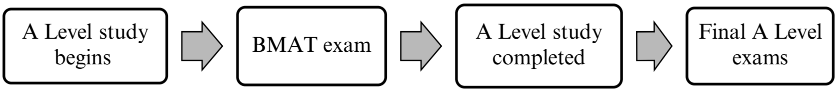
> The logical distinction between predictive validity and concurrent validation is based not on time, but on the objectives of testing. Concurrent validation is relevant to tests employed for diagnosis of existing situations, rather than prediction of future outcomes. The difference can be illustrated by asking "Does Smith qualify as a satisfactory pilot" or "Does Smith have the prerequisites to become a satisfactory pilot?" The first question calls for concurrent validation; the second for predictive validation (Anastasi and Urbina 1997:119).

Sometimes it is easy to distinguish between concurrent and predictive validity. When a criterion is measured well into the future from the point of an assessment, and is intended to provide evidence that an assessment can meaningfully predict future outcomes, it is clearly predictive validity. This is the case when admissions tests are correlated with measures of academic performance at university. However, in other cases it can be very difficult, and potentially unimportant, to distinguish between concurrent and predictive validity.

One confusing issue is that the word 'prediction' is used in two ways as explained by Anastasi and Urbina (1997:119): 'criterion-prediction validation procedures indicate the effectiveness of a test in predicting an individual's performance in specified activities . . . the term 'prediction' can be used in the broader sense, to refer to prediction of the test to any criterion situation, or in the more limited sense of prediction over a time interval'. Therefore, a statement that 'assessment X predicts Y' may refer to a statistical correlation that is either concurrent or predictive. In this volume, we use 'predict' in the broader statistical sense, to describe a relationship between two variables.

To illustrate the challenge in distinguishing between concurrent and predictive validity, let us consider research designs that correlate A Level grades (a measure of academic achievement) with BMAT scores (see Figure 6.2).

Figure 6.2 Typical A Level and BMAT arrangements in the UK



Studying for school examinations tends to begin in advance of sitting a university admissions test, and awarding for some components that contribute to the final grade can occur early in the course of further education study. For qualifications or assessments that are composed of various subcomponents awarded throughout a course of study, the question of when it was administered is not a simple one to address. On the other hand, final grades for school-leaving qualifications might not be available until after the results of the admissions test are released; this makes it reasonable to treat A Level grades as criteria predicted by scores on the admissions test.

Research on A Level performance and BMAT scores can be regarded as examples of concurrent or predictive validity studies, depending on the intended use of BMAT scores in the selection process. Due to the difficulties in timing as explained above, we do not describe research investigating school qualifications and test scores as concurrent or predictive validity. Instead, those using the research can determine whether they would consider the studies as concurrent validity, or predictive validity, or whether to consider the results more generally as evidence of criterion validity, which may be informed by their intended interpretation of the findings.

Exploring these relationships is still important, because it is desirable that BMAT scores should relate to other measures of potential for biomedical study. However, criteria used in student selection should aim to have *incremental* predictive validity over other criteria, contributing some unique information on applicants' potential. For example, we would expect BMAT scores to be related to academic measures such as A Level attainment, which are known predictors of future course performance. On the other hand, a high degree of shared variance might imply redundancy of measures (unless there was doubt about the predictive equity of A Level grades for different groups of applicants). Given that the selection process generally begins before final A Level grades are available for most university applicants, a positive relationship between BMAT scores and A Level attainment (as an outcome criterion) will prove useful to selecting institutions. An example of how BMAT scores have been shown to predict both high and, importantly, insufficient A Level attainment (the failure to achieve minimum conditional offer grades and therefore face a late rejection) is presented later in this chapter (Emery 2007c). In line with Anastasi and Urbina's (1997) focus on distinguishing between concurrent and predictive validity based on the intended use of scores, these findings can be interpreted as either aspect of criterion-related validity.

Box 6.3 BMAT scores' correlation with A Level grades

Tip: BMAT scores have been shown to correlate with A Level grades. This can be useful to schools making admissions decisions before A Level grades are known.

One other aspect of concurrent validity that has not been investigated is the relationship between scores achieved on two different versions of BMAT. Administering two test versions of a test to the same group of students can be used to establish parallel forms of reliability (see Chapter 5 for a description of this), and to investigate the equivalence of writing tasks in BMAT Section 3. Conducting these studies in the future will provide concurrent validity evidence for BMAT.

Methodological challenges

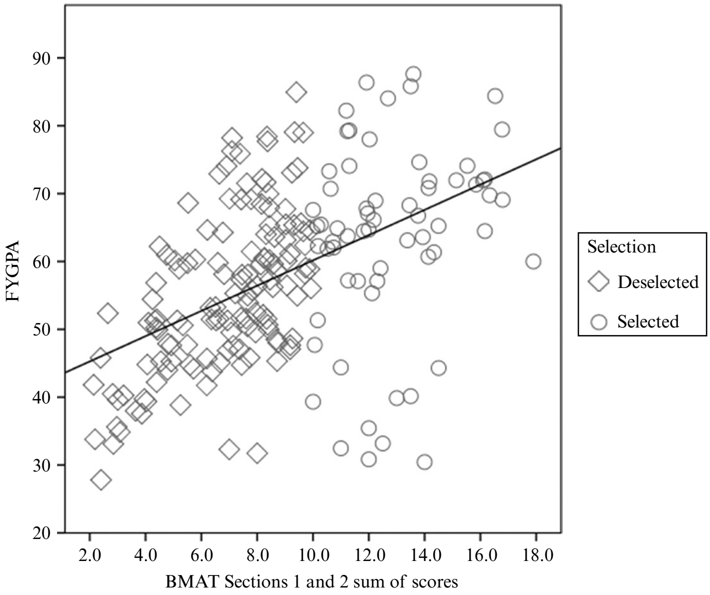
Establishing predictive validity evidence is a priority for tests like BMAT as relevant outcome data becomes available for the test takers. This evidence is important to stakeholder institutions and to test takers themselves, so predictive validity formed the focus of much early research work on BMAT. However, it is difficult to establish predictive validity evidence for university selection tests because methodological issues systematically reduce the strength of correlations that are observed in datasets used for investigating predictive validity. In this part of the chapter, we present three of the main ways that correlations are attenuated in selection contexts, using simulated example datasets. Following an overview of these issues, the approach to presenting predictive validity analysis adopted by Cambridge Assessment researchers is outlined.

Range restriction

Range restriction arises because predictive validity must be calculated from the pool of accepted applicants, whose test scores represent a selected range that is higher and narrower than that of the overall applicant pool. The course performance of applicants who were rejected with low test scores cannot be known and researchers are restricted to looking for differences in course performance between selected applicants, who typically achieved in the range of scores deemed adequate by selectors. Test scores on BMAT can be used in a variety of ways and this can impact substantially on the relationships observed in analyses. The rejection of low scorers, particularly if there is a minimum score that applicants must achieve to be accepted onto the course, gives rise to restricted ranges of test scores and shapes of scatter that limit the strength of relationships. A detailed description of various types of range restriction is available in Bell (2007), but for the purposes of explaining the concept of range restriction, a single hypothetical example will suffice. Low scorers can be rejected by setting a minimum score, which is often referred to as applying a cut-score or hurdle; this method is commonly used in selection settings and results in the forms of range restriction that are easiest to conceptualise.

Consider an idealised situation where the sum of an applicant's scores on BMAT Section 1 and 2 is correlated with FYGPA at $r = 0.492$ (see Figure 6.3). Of course, applicants to the course that are located in the bottom left of the plot are unlikely to be accepted onto the course; therefore, their FYGPA would not actually be observed.

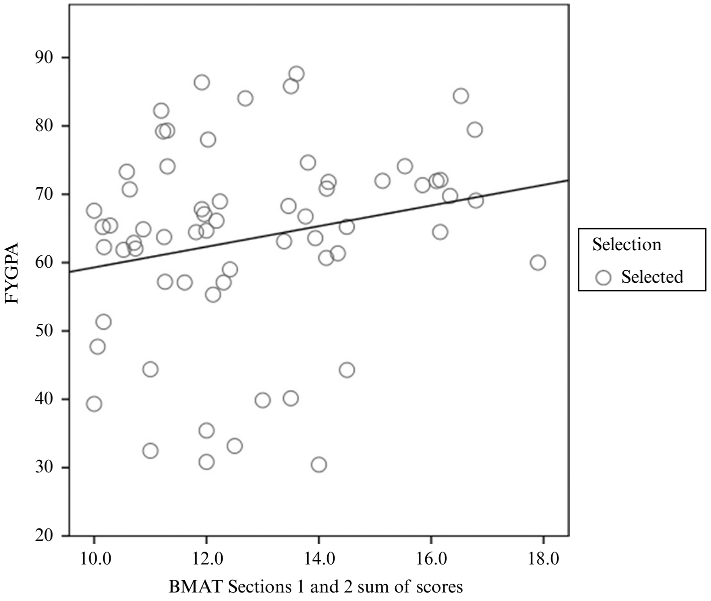
Figure 6.3 Idealised correlation between BMAT scores and FYGPA



If the admissions process applied a hurdle so that only applicants with a combined BMAT Section 1 and 2 score of 10 or above were accepted onto the course, only those applicants on the right-hand side of Figure 6.3 would have course performance data available for analysis, resulting in Figure 6.4. When the correlation between the two variables is calculated only using the data observed after applying a hurdle, the coefficient indicates a much weaker relationship of $r = 0.218$. As the correlation coefficient is an estimate of the relationship that exists between the two variables, the observation of a weaker relationship is described as an attenuation of the estimate. The observed statistic is much weaker than the relationship that would be observed if the data in the entire population was available.

This simplified example demonstrates one of the major challenges with estimating the relationship between a variable used for selection and an outcome score. While use of a cut-score for accepting applicants onto a course may be appropriate for some selection contexts as an admissions practice, it can exacerbate problems in assessing predictive validity. One method for dealing with this issue is to apply formulae that correct the strength of the coefficient based on the distributions of the variables (e.g. Sackett and

Figure 6.4 Correlation between BMAT score and FYGPA after selection using a cut-score on BMAT



Yang 2000). This has been used in research on admissions tests by McManus, Dewberry, Nicholson, Dowell et al (2013) and represents a suitable solution for the example we have outlined. However, we should recognise that the situation represented in Figure 6.3 and Figure 6.4 is a simplistic one. It is far more common for biomedical courses to use multiple stages in their selection process. This means there might be hurdles applied on various selection criteria that restrict the range of observed scores in different ways, either directly or indirectly. Furthermore, predictive studies often use more than one outcome as criteria, which can complicate things even more. In some instances, collegiate systems used by Oxford and Cambridge can even mean that subgroups within a course cohort had different hurdles applied to the selection criteria that were used, possibly in a different order. Without a detailed understanding of the mechanisms used in selection, it can be very difficult to unpick the ways that a final cohort of students was derived. In general, the greater the reliance on a test score in selection, the more that range restriction becomes an issue when calculating correlations.

Compensatory selection

Within a multi-faceted admissions process, poor performance on one selection criterion (an admissions test) might be compensated for by good

Figure 6.5 Idealised correlation between BMAT scores and FYGPA indicating outlier scores

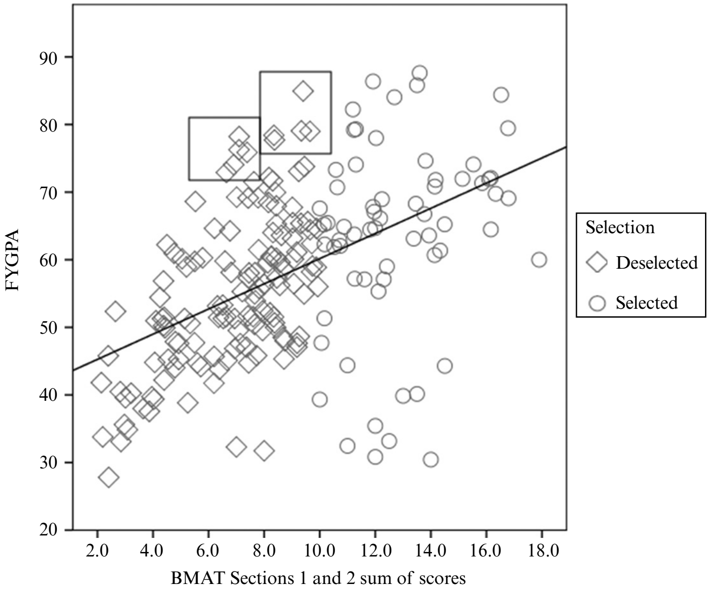
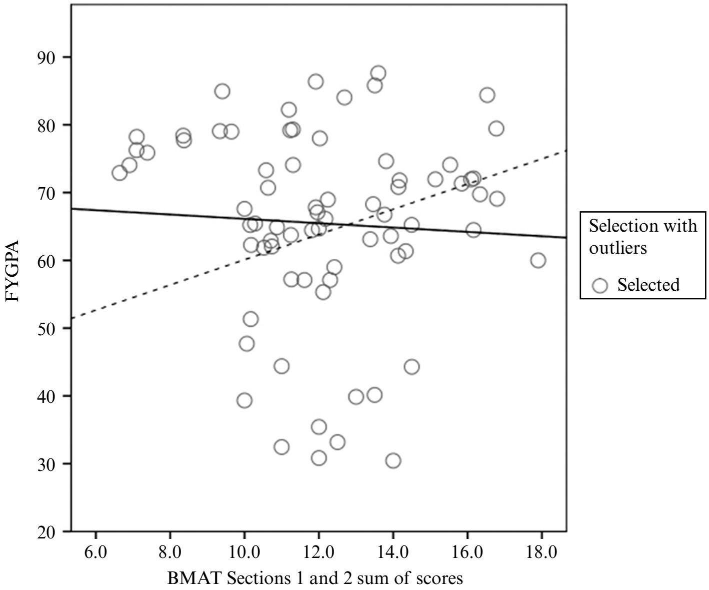


Figure 6.6 Correlation between BMAT score and FYGPA when the candidates with outlier BMAT scores are also selected



performance on another (e.g. academic achievement, interview). While an appropriate practice for university admissions, this again creates problems for analysing predictive validity. The compensatory use of assessments in multi-method selection procedures means those accepted onto a course with low test scores tend to have performed well on other selection measures. These candidates are therefore atypical of low test scorers in general terms of their potential for succeeding on the course. Compensation in selection processes can make predictive relationships appear, statistically, to be non-existent or even negative. In order to illustrate the effects of compensatory selection on predictive relationships, let us reconsider the hypothetical scenario described previously. Figure 6.5 depicts the idealised correlation between BMAT combined scores and FYGPA as before, but several outlier scores have been highlighted by the boxes, which represent candidates who achieve just under the hurdle of a combined BMAT Section 1 and 2 score of 10. Such candidates may be selected if they have strong performances on other admissions criteria (e.g. school grades or interview scores) that indicate they will be successful on the course. In other words, high performance on some selection criteria can be used to compensate for lower performance in another, allowing the admissions tutor to identify applicants more likely to succeed on the course, when compared with other applicants who achieved the same test score.

Figure 6.6 shows the resultant correlation between BMAT scores and FYGPA when these candidates are selected alongside those above the hurdle. As we can see, the observed correlation, represented by the solid line, is weak and negative (r = -0.058). Inclusion of the outliers has rendered the population correlation, represented by the broken line, undetectable. This is an extreme example because it assumes that the admissions tutor is very accurately identifying those applicants who do not meet the test score threshold but are likely to do well on the course. However, it illustrates how selecting a small number of applicants who are atypical for their test scores can impact on criterion validity.

Again, it should be recognised that the situation represented in Figure 6.5 and Figure 6.6 is more simplistic than actual selection contexts. It is common for biomedical courses to consider multiple selection criteria in their procedures, which may be weighted in various ways. Furthermore, qualitative data and contextual information such as indicators of low socio-economic status are often considered alongside other selection criteria.

Confounds on the outcome variable selected as a criterion

Predictive validity is also affected by potential confounds on the outcome variables. For example, academic outcomes in university can be affected by teaching quality and practice over a course of study. Similarly, outcome measures can be affected by unreliability in course assessments, for example, introduced by subjective marking of assignments. The potential for grades to be confounded as an outcome variable is described by Stemler:

> Grades can be difficult to interpret because they are so frequently influenced in non-uniform ways by other factors. For example, not everyone goes to the same university. Not everyone in the same university takes the same courses. Not everyone in the same courses has the same instructor. Sometimes the interpersonal relationship (either good or bad) that a student has with an instructor colors the instructor's evaluation of the student's content mastery. Each of these outside factors, and many more, can influence final course grades in ways that are not always related to the cognitive abilities and traits that reside within the student (Stemler 2012:8).

The example given by Stemler illustrates the issues that may arise when aggregating grades to serve as outcome data across institutions, courses or years. Collapsing data can produce outcome variables that vary in ways not related to the construct of interest.

In medical education, courses within a university tend to be consistent with all students undertaking a relatively homogenous curriculum of study. Similarity in the course of study undertaken by students reduces the potential for confounds on GPA as an outcome criterion. One caveat to this is that biomedical courses typically include optional components that are selected by the students. Thus, cumulative GPA may be made up of different sets of marks across subgroups of students. This can mean that GPA appears to be comparable across all students, when it is actually composed of differing proportions of assessment types, and actually represents a mixture of subconstructs. In addition, selection of optional components is unlikely to be random, and may have a systematic relationship with abilities assessed as part of selection. For example, those students confident in their mathematics abilities are more likely to select study options with substantial quantitative components. Further down the line, choices about the medical specialty to pursue might be linked to the communication skills that were assessed during interviews. All of these factors can reduce the reliability of GPA as an outcome measure in predictive validity analysis. However, while we acknowledge the limitations in using GPA in predictive studies, we maintain that within biomedical programmes it is a valid outcome criterion. On this issue, other admissions testing researchers agree, and Stemler writes: 'It is perhaps the case that [first year] GPA is the best that we can hope for as a proxy of domain-specific knowledge at this time' (Stemler 2012:8).

In conducting predictive research, it is important to recognise the limitations of measures that represent academic achievement in university, and to interpret findings whilst considering these issues. We should not assume that everyone's experience when studying will be the same. Indeed, there can be phenomena that change the learning context systematically in line with the constructs being assessed. This would reduce observed relationships between a selection measure and outcome measures. For example, a medical or dental school might identify those candidates with deficits in their scientific knowledge (using the selection test or early indicators of performance on the course). The department can bring this to the attention of the student, increasing the likelihood that the student focuses on this area more than their fellow students. In some instances, the university may offer extra support to students. This is particularly common with written communication skills where students are referred to central support services if their writing is not good enough. If these interventions and influences on behaviour are effective, they can mask the relationships between performance on the selection measure and course outcomes.

Finally, it should be noted that motivation has a large part to play in students' learning and study behaviour, and has been linked to course performance outcomes (Kusurkar, Ten Cate, van Asperen and Croiset 2011). Although biomedical students are typically assumed to be highly motivated, motivation is not considered a stable characteristic and is dependent on contextual factors (Pelaccia and Viau 2017). For example, Wouters, Croiset, Galindo-Garre and Kusurkar (2016) showed that the motivation of applicants to study medicine was high immediately after selection, but decreased rapidly after entering medical school. Thus, changes in motivation over a course of study or even during the transition from secondary to higher education may weaken the relationship between performance on a selection test and outcome measures.

In summary, course performance is affected by a great many variables beyond what the test aims to measure. Course performance will be influenced by a multitude of factors, including the educational environment and the personal circumstances of a student, in addition to academic potential as measured through an admissions test. Even variables traditionally considered as stable, such as personality traits, are now acknowledged to include some plasticity and adaptability (Ferguson and Lievens 2017). Therefore, it is important to recognise that situational context will account for some variances in learning behaviours, which will also impact on course outcomes. The moderating effects of factors such as conscientiousness, motivation and non-academic issues on the relationship between academic ability and course performance should not be underestimated given the amount of variance that remains unexplained by academic ability measures. A single outlying data point from a candidate who has a good test score but fails the course for unrelated reasons can have a large impact on the size of a correlation coefficient. For these reasons, the magnitude of correlation coefficients regarded as beneficial between selection tests and future performance are lower than would normally be expected. In addition, the potential for confounds to reduce the reliability of the criterion increases with the length of time between the test administration and the measuring of the outcome being predicted.

The Cambridge Assessment approach to reporting predictive validity

When describing the methodological issues that result in weaker observed correlation coefficients, we have stressed that the examples used are simplified versions of how selection functions in practice. The ways that actual procedures influence the findings of predictive research are complex and easily overlooked. In fact, the example used to illustrate range restriction actually included a compensatory selection method, even if the hurdle were applied strictly without considering other selection variables. This is simply because two separate BMAT sections were combined to form an aggregate score. Applying a hurdle in this case means that scores on the two sections effectively compensate for each other, so that the hurdle can be reached by a low score on one section with a higher score on the other. Whilst this might be suitable for an admissions policy, it means that relationships between the two section scores would be weaker in the selected group than they are in the wider pool of test takers. As the section scores are expected to predict specific course components differently, the interactions can have further impact on the relationships observed in predictive validity studies.

For selection tests, valuable criterion validity evidence can be obtained from a pilot year of the test where selectors are blind to applicants' scores. The use of such a pilot year can help overcome issues of range restriction and compensatory selection. If it can be shown that candidates scoring low on the test have a very low chance of being offered a place of study following interview when scores were unseen by selectors (the criterion being the admissions decision) then this provides justification for setting a cut-score on the test in future years. Concurrent validity would be evidenced by high agreement between the (hypothetical) admissions decisions made by test scores and the admissions decisions made by selectors blind to those test scores, perhaps on the basis of interviews. Such evidence would allow an institution to apply a future cut-score on the test as a hurdle to the interview stage so as to focus their interview resources on applicants with a reasonable chance of gaining a place of study.

It is not always practical to pilot an admissions test and ignore test scores when making selection decisions. Admissions tests are often introduced in response to specific logistical issues, such as heavy oversubscription or ceilings in applicants' school-leaving qualifications. In these contexts, it might be necessary to use test scores immediately in some way. Even if a hurdle is not applied, it can be difficult to justify the logistical demands of administering a test if admissions tutors do not have access to test scores. Additionally, there are ethical implications of requiring applicants to sit an exam where the scores will not be considered. This means that studies of criterion-related validity are often conducted in the context of real-world selection, requiring researchers to consider how the pool of accepted students has been shaped by admissions decisions.

Given the complexity of the selection contexts we have outlined, Cambridge Assessment recommends that uncorrected correlation coefficients are reported for predictive validity studies, alongside an account of the decisions that resulted in the student cohort that was selected. In order to acknowledge the tendency for coefficients to be attenuated by various issues, guidelines for interpreting uncorrected coefficients can also be included with results, such as those in Table 6.1, which were published by the US Department of Labor, Employment and Training Administration (1999). Alternatively, Cleland et al (2012) provide similar principles for interpreting correlations in the medical admissions context.

Table 6.1 Guidelines for interpreting correlation coefficients in predictive validity studies

Validity coefficient	Interpretation
Above 0.35	Very beneficial
0.21 to 0.35	Likely to be useful
0.11 to 0.20	Depends on circumstances
Below 0.11	Unlikely to be useful

Source: US Department of Labor, Employment and Training Administration (1999)

As with all rules of thumb used to interpret statistical analyses, these values should not be interpreted blindly. It is important to try and obtain as many details relating to the selection procedure used as possible.

One final point that can help deal with the methodological challenges described here is not actually part of our approach, but more of a lesson learned from conducting predictive studies. Most, if not all, of the predictive validity work conducted by Cambridge Assessment has been retrospective. In other words, the studies have been designed and conducted once course outcomes became available, by accessing historical records. From our experience, retroactively describing all of the selection decisions made in application cycles can be difficult to complete with precision. Therefore, it can be advantageous to plan predictive studies prospectively, as this enables the selection decisions used at various stages to be documented in their entirety. In future, we intend to plan predictive research prospectively, as this could allow statistical corrections to be used on observed correlations with confidence.

Box 6.4 Key recommendations from the Cambridge Assessment approach to predictive validity

- Select outcome criteria that are theoretically relevant to the test construct.
- Criterion validity is best established during a 'pilot' year in which the test is administered but admissions decisions are made 'blind' to test results.
- Uncorrected correlation coefficients can be reported alongside descriptions of the selection context and guidelines for interpreting attenuated coefficients.
- It can be useful to plan predictive studies prospectively.

Collecting and collating data

Conducting research on criterion-related validity is dependent on collecting data that is external to the test scores themselves. However, there are difficulties with collating datasets with sufficient sample sizes, as selected cohorts for biomedical courses tend to be small.

Some researchers advocate conducting multiple site or multiple cohort studies to increase statistical power (e.g. McManus, Dewberry, Nicholson and Dowell 2013). A recent big data initiative will enable large-scale predictive validity studies to be conducted more easily in the UK, by collating data from test providers, medical schools and the royal medical colleges into a UK Medical Education Database (UKMED). Big data can offer opportunities to better understand the factors that contribute to success in medical study, and UKMED seeks to support large-scale medical education research by allowing researchers to combine and analyse anonymised datasets. This will encourage studies with large sample sizes and greater statistical power.

The UKMED project is managed by the medical regulator in the UK, the General Medical Council (GMC), and Cambridge Assessment is in the final stages of contributing BMAT data to the database, whilst considering some of the data privacy concerns raised by commentators (e.g. Best, Walsh, Harris and Wilson 2016). The larger scale studies enabled by big data approaches are useful for investigating how the predictive validity of an assessment might generalise across different contexts; however, there are several issues with research using data from different institutions across multiple years. Firstly, adopting Hopkins et al's (1990) approach of treating any variable that correlates with performance as a valid selection criterion could lead to selection methods with no theoretical basis being used, which would also result in unintended side effects for the professional workforce.

To safeguard against identifying spurious relationships, we advocate cautious use of large databases by relying on theory to develop hypotheses about expected relationships. Formulation of relevant hypotheses should include consideration of consequential validity and test taker characteristics. For the UKMED project, research proposals are scrutinised by a group of researchers, which includes a member of the Cambridge Assessment research team that focuses on admissions testing research. Furthermore, Cambridge Assessment also participates in UKMED's Advisory Board, which is the governance structure for the project.

Other methodological issues result from difficulties in obtaining the relevant information about how procedures were applied. If precise details about different selection practices (e.g. hurdles and compensatory selection methods) are not known, it can be difficult to validly adjust for them. As mentioned earlier, this presents complicated issues when considering a single course of study, so documenting the impact of these issues across multiple courses can be even more complex. Furthermore, biomedical courses differ in their composition, affecting the comparability of outcomes across courses, or even within courses that have optional components. The impact of this variability can be reduced by standardising indicators of course performance within cohorts before including them in statistical models, but this does not entirely mitigate the issues faced when combining data across consecutive years of study.

Moreover, different courses, or even course components, can have varying relationships with test sections or selection methods, due to differing candidatures or content focus. For example, one course might have more components that focus on natural sciences than another course that has more assessments that rely on written communication. Performance in these courses would theoretically have different relationships with BMAT Sections 2 and 3. Treating both courses as the same by including them in one analysis can mask nuanced relationships between course outcomes and selection criteria with different emphases, which might be more easily detected using separate analyses.

Therefore, smaller scale studies can contribute effectively to establishing the predictive validity of admissions tests and should not be automatically overlooked in favour of studies with greater statistical power. The approach adopted by Cambridge Assessment researchers in this regard is to collaborate with admissions tutors at universities using BMAT, in order to support them with their own evaluations of predictive validity, which tend to be smaller scale than multi-site studies. This acknowledges that the test users are experts with substantial knowledge of the selection context, whereas Cambridge Assessment researchers tend to be more familiar with issues in educational assessment, such as the impact that various methodological challenges can have on statistical analyses. Most, if not all, BMAT users have conducted their own evaluations of the test, which typically include analysis of predictive validity. This allows individual departments to interpret results in the context of their own courses, in order to satisfactorily show that selection procedures are suitable to institutional committees.

There is often mutual sharing of BMAT data, admissions information, admissions decisions and course performance data between Cambridge Assessment and the institutions using BMAT, which allows both organisations to monitor the predictive validity of BMAT for admitted applicants. These studies provide valuable insights into BMAT's validity; however, Cambridge Assessment researchers arrange to analyse data on a case-by-case basis for each study. As personal data is often required to match course data to BMAT scores, we review data protection issues separately. The on-course performance of students who have taken BMAT is not routinely collected from universities in the same way as some other test providers; this reflects a cautious approach to data protection throughout the Cambridge Assessment Group, which is informed by recent discussions on the opportunities and risks presented by use of student data (e.g. Trainor 2015).

Box 6.5 Key points on data collection

- Small-scale studies can provide important findings that complement large-scale studies.
- Each course or institution's selection procedure is unique and test developers can collaborate with test users to investigate predictive validity.
- BMAT data is being included in a database managed by the GMC, which will support large-scale research into the validity of selection criteria.
- Cambridge Assessment collaborates with universities on a case-by-case basis, and does not routinely collect large amounts of data on candidates from biomedical departments.

Data on selection criteria (e.g. A Level results) is available via the Universities and Colleges Admissions Service (UCAS) to biomedical and dentistry schools for all university applicants rather than just those admitted, as is candidate-level demographic information. This can permit research into institutions' selection processes in general, such as the fairness of admissions offers for different candidate groups, the relationship between demographic variables and selection criteria, and the factors that best predict admissions offers. Predictive validity studies on BMAT are typically single-institution studies using one or more cohorts, which helps reduce confounds on the outcome variable (as described by Stemler 2012). While cohorts tend to be analysed separately, it is sometimes necessary to combine them across years for courses which have particularly small numbers, such as graduate-entry courses (Devine and Gallacher 2017). Findings from separate analyses often illustrate the variability in the strength of correlations that can be found even within the same course, resulting from different admissions decisions, applicant cohorts or course assessments. Outcome data usually consists of early course examination results (e.g. end of Year 1 or Year 2 examination average).

Predictive equity and its role in test fairness

An aspect of predictive validity research that is crucial to investigating test fairness is that of *predictive equity*. This is discussed in detail in Chapter 2, and is illustrated by research into test fairness and bias issues (Emery et al 2011). To recap here, if a test is biased *against* a particular candidate group then we would expect test scores to systematically *under*-predict future course performance for that group (i.e. they go on to perform better than predicted on the course), and vice versa. If BMAT scores fairly reflect ability on the construct of interest regardless of candidate group then scores should predict future course performance equitably for different groups, assuming that other factors are equal between them.

Candidate school sector information and candidate gender are therefore included as additional predictor variables in Cambridge Assessment Admissions Testing regression analyses of course performance on BMAT scores and it is possible to investigate any other candidate-level variables that may be a fairness concern. If a given BMAT score predicts equal course performance, on average, between different candidate groups then this provides strong evidence that the test is fair and unbiased even when test score differences are evident between groups. Analyses to date have consistently shown BMAT to predict course performance equitably regardless of candidate background variables such as gender, school type, school sector and social deprivation indicators.

6.3 Research

In the previous parts of this chapter, we have described the theoretical and methodological issues involved in conducting criterion-related validity, and the approaches to addressing these that are used in BMAT research. Research into the predictive validity of BMAT is regularly conducted by Cambridge Assessment Admissions Testing in collaboration with the universities who use the test. In this section, we present a longitudinal study – conducted when BMAT was first introduced at Cambridge Medical School – which provided foundational evidence of BMAT's predictive validity to support its use as an admissions test for medical study (Emery and Bell 2009). We then summarise some of the other predictive validity studies which have been conducted, with a focus on the diverse contexts that BMAT is used in.

Key research study – The predictive validity of BMAT for pre-clinical examination performance (Emery and Bell 2009)

Main findings

- BMAT makes a significant contribution to predicting performance in medical study.
- BMAT makes a unique contribution to predicting performance when considered alongside other selection criteria.
- Section 2 correlated most strongly with performance in pre-clinical courses.

Introduction and context

The following study was one of the earliest pieces of predictive validity research carried out with BMAT, and provided foundational evidence for use of BMAT in medical student selection (Emery and Bell 2009). This investigated the predictive validity of BMAT (and its predecessor, Medical and Veterinary Admissions Test (MVAT)) in the first four years of use as a selection tool at the medical school of the University of Cambridge. Outcome variables investigated were first and second year medical school performance (both examination marks and examination classification) in four individual cohorts of students.

BMAT was introduced in order to address several problems that University of Cambridge's medical school was facing. Firstly, the applicant pool comprised students with very high, but similar levels of prior academic achievement (A Level grades or equivalent), making it difficult to distinguish between applicants. In addition, there were other problems with reliance on prior school achievement as a selection criterion, such as the need to consider non-UK applicants, the attainment advantage of those attending private schools, the poorer performance of various social groups and the fact that only predicted A Level grades are available at the time of application. BMAT was used as an *additional* source of information to help selectors differentiate between those with the high prior attainment and to compare students from different educational backgrounds and countries.

With such strong competition for places, it is important to establish that a selection measure has predictive validity if test takers and institutions are to have faith in its fitness for purpose. The aim of this study was therefore to determine whether BMAT scores were a significant predictor of early medicine course performance (science-based examinations) in four cohorts of students who were all admitted with the highest A Level grades possible at the time. If selection test scores can significantly predict course performance in students admitted with uniformly high A Level grades (or significantly predict course performance after controlling for A Level grades) then they are a useful addition to the selection process and will be beneficial in increasing student success rates (Kuncel, Hezlett and Ones 2001). The magnitude of the predictive relationships and their variability over course components and cohorts was investigated, given that this appeared to be a typical finding elsewhere (Julian 2005). Whether BMAT Section 1 or Section 2 showed the stronger predictive relationship with course examinations was also of interest.

Research questions

1. Does BMAT significantly predict end of Year 1 and Year 2 examination performance in four cohorts of students entering the medicine course at the University of Cambridge[5]?
2. What is the relative magnitude of the predictive relationship for Sections 1 and 2 of the test?

Data collection and analysis

The medicine course at the University of Cambridge is a 'traditional' (rather than 'integrated') course in that it consists of three years of pre-clinical study followed by three years of clinical training. The first two years are heavily science based. For the cohorts in this research, students completed three core first year courses and four core second year courses, each assessed in examinations at the end of the academic year. The examinations each consisted of a mixture of short-answer, essay and multiple-choice questions based on lecture and practical material. Third year outcome data was pass/fail in nature (and composed of a large number of course options not necessarily related to medicine) and so was not included in the study. Pre-clinical course examinations were the focus of this study as BMAT focuses on academic readiness for demanding science-based study and not clinical skills/fitness to practice.

Scores for Sections 1 and 2 of the test correlated at around 0.4 in these cohorts (as they do in general). It should be noted that Cambridge Assessment did not mark Section 3 prior to 2004 and the University of Cambridge did not use BMAT Section 3 (Writing Task) scores in selection in these test years (2000–03), instead considering candidates' responses as a qualitative piece of evidence and to promote discussion during the interview. Thus, Section 3 scores were not analysed in this study. No BMAT cut-score was applied as a hurdle to the interview stage for these cohorts of applicants, meaning that a full range of BMAT scores was technically possible.

[5] BMAT was known as MVAT prior to 2003 so the first three cohorts in this study sat MVAT rather than BMAT.

Examination data for the first and second years of the medicine course was supplied by the University of Cambridge and matched to students' MVAT/BMAT results. Examination data consisted of a total (percentage) mark for each course component plus an overall (percentage) mark and examination classification for each year. First year examination classes were, in descending order of merit: 1st, 2nd, 3rd, Fail. Second year examination classes further subdivided the 2nd class into higher and lower categories. Attrition rates are very low at this institution and numbers were too small to permit its analysis for these cohorts. Around one fifth of each first year cohort and one sixth of each second year cohort was awarded a 1st class.

Numbers of students with complete data in each cohort were 255, 250, 247 and 250, respectively. A small number of students in each cohort could not be matched to MVAT/BMAT results. Fewer than 10 students in each cohort were aged over 21 at the time of course entry. Students gave permission for their examination and test scores to be used for research purposes when registering for the test and data was anonymised after matching. The four cohorts were analysed separately.

Pearson correlations were employed with the examination marks data, which were continuous and normally distributed. Upward adjustments of the correlation coefficients for range restriction were not applied because the complexity of the selection process, a compensatory mixture of qualitative and quantitative information, made them inappropriate (Sackett and Yang 2000). Raw, uncorrected correlation coefficients were therefore presented throughout the results. Logistic regression analyses were employed with the examination classifications in each year, modelling the probability of achieving a 1st class result as a function of BMAT Section 1 and BMAT Section 2 scores.

A Level grades could not be included as an additional predictor variable in this study as there was a ceiling in grades for the cohorts included. A Level grades AAA were required for course entry for these cohorts, which was the maximum attainable outcome at the time (prior to the introduction of the A* grade in 2010).

Results

The score distributions of those offered a place versus those rejected in each cohort shows that those who received an offer had a higher mean and narrower range of test scores than those who were rejected but there was considerable overlap in their distributions. A number of applicants with relatively low test scores were offered a place each year and a number of high scorers rejected due to the compensatory nature of the selection process.

1. Correlations with course examination marks

Table 6.2 displays the Pearson correlation coefficients between MVAT/ BMAT scores and the Year 1 and 2 examination marks. It can be seen that the strength of the relationships varied across the cohorts and course components but they were consistently stronger for Section 2 of the test (Scientific Knowledge and Applications) than for Section 1 (Aptitude and Skills) in these students. The majority of coefficients for Section 2 fell within the 'very beneficial' range (above 0.35) or the 'likely to be useful' range (above 0.21). Correlation coefficients were slightly weaker for the second year examinations, an outcome expected given that predictive relationships typically weaken with increasing time intervals (Julian 2005). The exception was Section 1 for the BMAT 2003 cohort, which correlated more strongly with their second year examination performance.

Table. 6.2 Pearson correlation coefficients between BMAT scores and examination performance

Section 1 – Aptitude and Skills

	Year 1 examination components			
Cohort	**Homeostasis**	**Molecules in medical science**	**Functional architecture of the body**	**Total mark**
MVAT 2000	0.22***	0.27***	0.19***	0.24***
MVAT 2001	0.19**	0.17**	0.12**	0.18**
MVAT 2002	0.18**	0.22***	0.14*	0.19***
BMAT 2003	0.1	0.12*	0.11*	0.13*

	Year 2 examination components				
	Biology of disease	**Human reproduction**	**Neurobiology and human behaviour**	**Mechanisms of drug action**	**Total mark**
MVAT 2000	0.15**	0.13*	0.18**	0.24***	0.17**
MVAT 2001	0.12*	0.09	0.11	0.19***	0.11
MVAT 2002	0.20***	0.12*	0.04	0.22***	0.11
BMAT 2003	0.17**	0.20***	0.24***	0.16**	0.22***

Table. 6.2 (continued)

Section 2 – Scientific Knowledge

	Year l examination components			
Cohort	**Homeostasis**	**Molecules in medical science**	**Functional architecture of the body**	**Total mark**
MVAT 2000	0.45***	0.41***	0.40***	0.44***
MVAT 2001	0.28***	0.35***	0.26***	0.26***
MVAT 2002	0.46***	0.41***	0.41***	0.45***
BMAT 2003	0.28***	0.27***	0.16**	0.26***

	Year 2 examination components				
	Biology of disease	**Human reproduction**	**Neurobiology and human behaviour**	**Mechanisms of drug action**	**Total mark**
MVAT 2000	0.38***	0.34***	0.35***	0.36***	0.26***
MVAT 2001	0.29***	0.24***	0.24***	0.31***	0.18**
MVAT 2002	0.40***	0.17**	0.24***	0.42***	0.23***
BMAT 2003	0.23***	0.22***	0.27***	0.23***	0.25***

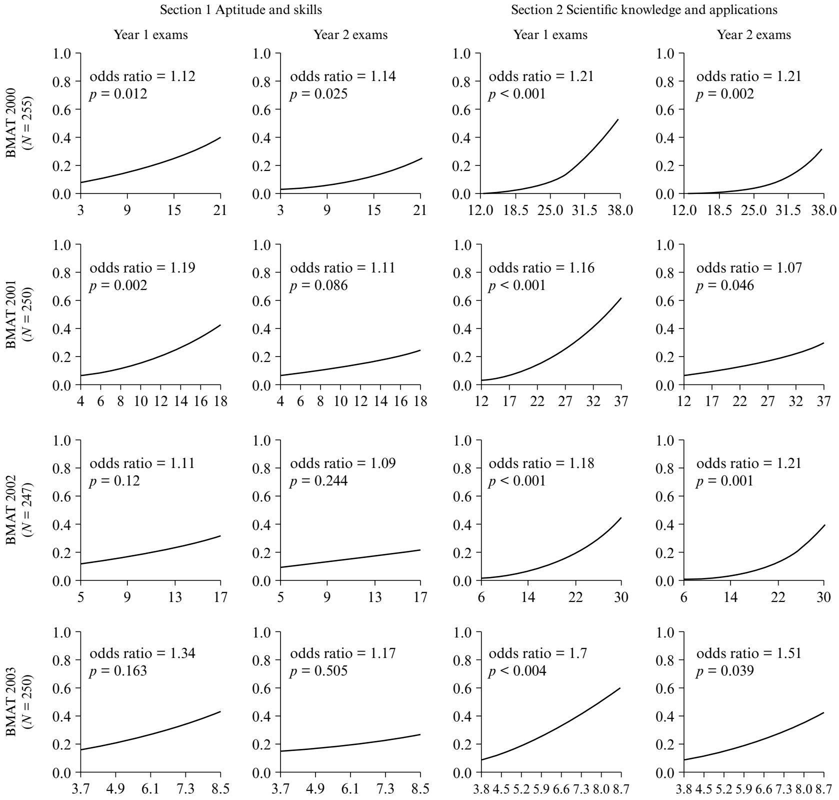
*Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$*

MVAT 2000 N = 255, MVAT 2001 N = 250, MVAT 2002 N = 247, BMAT 2003 N = 250

2. Prediction of high examination attainment (1st class)

The logistic regression plots in Figure 6.7 show the probability of achieving a 1st class examination outcome in Years 1 and 2 as a function of students' MVAT and BMAT scores. The x axes cover the actual range of scores achieved by each cohort. The steeper the curve, the stronger the predictive relationship (a horizontal function indicates no predictive relationship). Figure 6.7 shows that students' Section 2 (Scientific Knowledge and Applications) scores strongly predicted their probability of achieving a 1st class outcome in Year 1 and continued to significantly predict this in Year 2. Again, the plots suggest that relationship was stronger for Section 2 of the test in these cohorts (functions are consistently steeper than for Section 1). That is, an increase in Section 2 scores had the greater impact on the probability of achieving this outcome than did an increase in Section 1 scores. Note that the lowest Section 2 scores were associated with a very low probability of achieving a 1st class outcome in both years of the course. Odds ratios show the change in odds for every one point increase in scores on the x axes.

Figure 6.7 Logistic regression functions showing the probability of achieving a 1st class examination outcome in Years 1 and 2 as a function of MVAT/ BMAT Section 1 and 2 scores



Discussion

The results of this early study into the predictive validity of BMAT support the utility of the test for medical student selection. Correlations with examination marks compare favourably with those reported for the US Medical College Admission Test (MCAT), e.g. Julian (2005), particularly given that there is very little variability in prior attainment in this case because A Level grades were at a maximum in these four cohorts. This indicates that the test has incremental validity on top of prior academic achievement. Students who were accepted onto the course with low test scores, particularly on Section 2 (Scientific Knowledge and Applications), were unlikely to achieve the highest examination class. BMAT therefore appears to fulfil its purpose in identifying valid differences in the thinking skills and scientific reasoning of those with the highest possible A Level grades: differences that relate to future course performance (i.e. potential for biomedical study).

The correlations presented here are likely to be underestimates of the true predictive validity of BMAT. This is because correlations are attenuated for any criterion that counts towards selection due to the narrowing of score ranges (we cannot know how applicants rejected with low scores would have gone on to perform). Despite the lack of a cut-score at this institution and cautious use of the test in its earliest years, the scores of the accepted applicants showed a restricted range. The effects of compensatory selection on hampering the predictive relationship (i.e. the notion that accepted low scorers are likely to be atypically able) must also be kept in mind. The variation in the strength of correlations between cohorts even at the same institution is a typical finding (Julian 2005). For this reason, caution should always be exercised in citing a single number as a test's predictive validity coefficient.

Most of the correlations for Section 1 (Aptitude and Skills) were statistically significant but correlations were consistently stronger and logistic regression functions steeper for Section 2 (Scientific Knowledge and Applications) of the test in these cohorts. The findings from this early study are in line with many subsequent studies on the predictive validity of BMAT, which have also shown that Section 2 has stronger predictive validity. This finding also agrees with reported findings regarding the predictive validity of A Level chemistry and biology for early medicine course performance (e.g. McManus et al 2005). Stemler (2012) proposes predictive validity be tied to both domain-specific ability and domain-general ability. While Sections 1 and 3 in BMAT test domain-general ability (critical thinking skills and writing ability) and Section 2 assesses domain-specific ability (scientific reasoning), the criterion used for establishing predictive validity (course marks) is based on performance in pre-clinical courses, which is generally a measurement of domain-specific achievement. While the development of critical reasoning and problem solving skills is a common aim of medical education, a ubiquitous problem in establishing the predictive validity of critical thinking skills tests is that it is a domain-general ability that is rarely assessed within higher education (Stemler 2012).

Scientific reasoning with subject-specific knowledge (as assessed in BMAT Section 2) may predict course performance well because it additionally assesses motivation and interest in the area (Kuncel et al 2001, McManus et al 2005). A high BMAT Section 2 score suggests that a candidate thoroughly understands the scientific basics that underpin medical study and it is perhaps unsurprising that a poor score here is associated with a very low chance of obtaining the highest examination class.

It is widely accepted that there is much more to being a good doctor than academic success. However, success in science-based examinations is a necessary factor for progression to clinical training and a medical career regardless of whether it is sufficient for becoming a good doctor. It is the former and not the latter that BMAT aims to predict.

Summary of other relevant research

Cambridge Assessment has continued to investigate the predictive validity of BMAT for performance on medicine and veterinary medicine courses at the institutions using the test. This is particularly important to new institutions and courses adopting the test. In summary, the *strength* of the predictive relationship between BMAT scores and course performance varies between institutions, courses and cohorts. This variation may be explained by differences in how BMAT is used at different institutions, and aspects of the educational context that vary between and within courses. However, correlations are typically positive, with both Sections 1 and 2 of the test significantly predicting early course examination performance.

Results for Section 3 (Writing Task) are more mixed with regard to early course performance. In some cases, a relationship between BMAT Section 3 scores and indicators of course performance have not been observed, whereas in others Section 3 scores have been the strongest predictors of performance. Unsurprisingly, our findings indicate that Section 3 scores are more likely to correlate with modules assessed using written components, even where they do not correlate with overall course performance. This suggests it is useful to consider the content of course modules and how they are assessed, when interpreting observed relationships. Studies investigating criterion-related validity of BMAT in undergraduate courses, graduate-entry courses, and using A Level performance as a criterion are presented in the next sections.

Undergraduate course performance

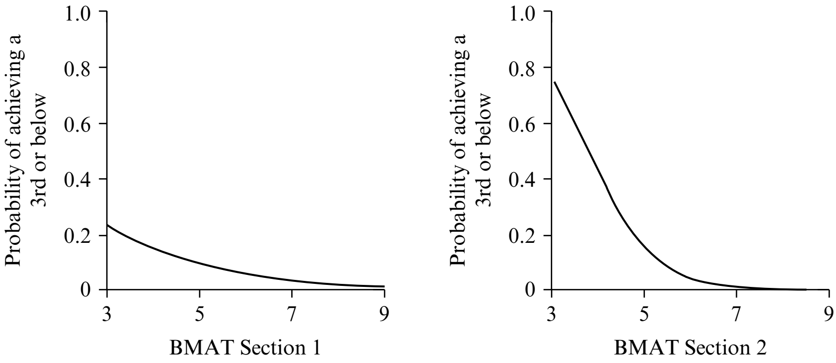
Whilst BMAT Section 2 tends to be the most consistently strong predictor of undergraduate course performance, this is not the case at all institutions and courses, or for all cohorts. For instance, Emery (2007a) showed an equally strong predictive relationship for both Sections 1 and 2 of BMAT for the University of Cambridge's 2004 veterinary medicine course but, in the following cohort (Emery 2007b), Section 2 was the stronger predictor of Year 1 course performance. As described in more detail below, Section 2 was not found to predict course performance in a graduate-entry medicine course (Devine and Gallacher 2017). In one institution, the Writing Task (Section 3) emerged as the strongest predictor of performance in two BMAT cohorts, with significant correlations in the range of 0.171 to 0.343 (e.g. Scorey 2009a).

Scorey (2009b) conducted predictive validity analysis of BMAT for undergraduate medicine performance at University College London (UCL), using successful applicants from the BMAT 2003–07 cohorts. BMAT had been used for selection purposes during these years, so direct range restriction was likely to have weakened the observed correlations. These analyses found that Section 3 scores significantly predicted early course performance (individual exam components as well as aggregate marks) in the 2004 and 2007 BMAT cohorts (correlations between 0.114 and 0.173). Moreover, Section 3 significantly predicted the probability of failing the second year of the course in the 2004 cohort; that is, as Section 3 scores increased, the probability of failing the course decreased. It should be noted that this study also revealed that Sections 1 and 2 predicted course performance in several BMAT cohorts and correlations tended to be stronger than the Section 3 correlations (correlations up to 0.254); however, there was some variation in correlation strength across all cohort and course year combinations.

Although such variation is a typical finding in predictive validity research, the reasons for these differences are difficult to establish. Differences in course content, teaching and examinations, and the ways scores are used in selection, will affect the nature of the predictive relationship but, for different cohorts on the same course, characteristics of the cohorts themselves or the way in which they were selected may be responsible for the observed differences.

Of particular note are the plots in Figure 6.8 (from Emery 2007b). In this cohort, unusually, there were sufficient numbers to permit analyses predicting a *poor* Year 1 medicine course examination outcome. In these analyses, students attaining a 3rd class result, failing the examinations or having left the course were categorised together (N = 20). It can be seen that, whilst BMAT Section 1 scores had little impact on the probability of this outcome (the x axes cover the score ranges of those who entered the course), students who were admitted onto the course with a low BMAT Section 2 score had a

Figure 6.8 Logistic regression functions showing the probability of achieving a poor examination outcome in Year 1 as a function of BMAT Section 1 and 2 scores (Emery 2007b)



high probability of a poor examination outcome. Those who were admitted with a BMAT Section 2 score of 5.0 or more had a very low probability of this poor outcome. This is an important finding given that all those admitted had achieved the maximum possible A Level grades in sciences prior to course entry.

Graduate-entry medicine

Cambridge Assessment Admissions Testing also investigates the predictive validity of BMAT for accelerated (graduate-entry) medicine course performance. More than a dozen UK universities offer accelerated (4- or 5-year) medicine courses for graduates with a degree in a scientific discipline; BMAT is currently used by three medical schools for graduate-entry selection. Graduate-entry courses also receive a large number of applications and the selection process is highly competitive. BMAT is a useful selection tool in the graduate-entry context because applicants have widely varying educational backgrounds and thus, are likely to have varying levels of foundational knowledge across the physical and biological sciences. An admissions test such as BMAT also allows broader access to graduate-entry medicine courses. For example, the additional information provided by BMAT enables admissions tutors to consider applicants from 'non-traditional' backgrounds, graduates from disciplines other than biosciences, and applicants who may be 'late developers' (i.e. with poorer A Level results but with a good degree classification). BMAT provides a common point of comparison between applicants from diverse backgrounds.

Recent analysis investigated the predictive validity of BMAT for graduate-entry medicine performance at the University of Oxford (Devine and Gallacher 2017), where shortlisting was done through grading of the applications by college and faculty tutors, with BMAT scores used only to differentiate candidates on the borderline of the shortlist; however, the score distributions indicated indirect range restriction of BMAT scores in the pool of shortlisted applicants. At Oxford, the preliminary examinations in medicine for graduates are made up of core and clinical examinations (awarded with a pass or fail), and five extension modules (awarded with percentage marks). In Devine and Gallacher's analysis, the five extension modules were included as outcome variables. Section 1 scores predicted average performance on the extension modules and correlated with performance on two course modules (correlations between 0.184 and 0.344). Section 3 (quality of content) scores were also found to correlate significantly positively with performance on the extension modules (correlations between 0.289 and 0.331). However, no significant correlations emerged between Section 2 scores and performance on the extension modules.

It is unclear why Section 2 scores did not correlate with performance on the extension modules in this cohort but it may be that knowledge of the secondary education level science curriculum has been replaced with more relevant biomedical knowledge from candidates' undergraduate degrees. That is, some graduate-entry applicants may not perform well on Section 2 if they can no longer recall the foundational knowledge, but may be able to learn medical knowledge more easily than expected due to knowledge gained during their undergraduate degree programme. This would reduce the strength of the relationship between Section 2 and course performance. Further work is needed to investigate this null finding, including analysis of performance on the core examination, which, due to its focus on basic facts and principles may have a stronger relationship with Section 2.

Nonetheless, the significant positive relationships between the other two BMAT sections and course performance suggest BMAT scores are likely to be useful for selection to graduate-entry medicine. In particular, the significant relationships identified between Section 3 scores and performances on the extension modules were encouraging, because scores from written essay tests have typically varied in their relationships with performance in medical study. For example, research looking at the essay component of the old (1992–2012) MCAT showed that writing section scores correlated only with some outcome variables, leading Hojat, Erdmann, Veloski, Nasca, Callahan, Julian and Peck (2000) to conclude that written communication skills are more closely associated with clinical practice than with achievement in the basic sciences.

A Level performance as a criterion

The criterion validity of BMAT with A Level outcomes has been investigated by Cambridge Assessment researchers (Emery 2007c). Given that only predicted A Level grades are available at the time of university application for most, selectors generally rely upon teachers' predictions for the majority of candidates. It is therefore of interest to stakeholder institutions if BMAT scores are correlated with outcomes at A Level, particularly to prevent places being offered to candidates who are unlikely to make the minimum grades required for entry. Whether BMAT scores correlate with two different A Level outcomes in a cohort of applicants (N = 460) was explored. These two outcomes were: the highest possible A Level outcome at the time of the study (grades AAA), and a poor outcome (failure to achieve the minimum offer grades of BBB). Correlations between BMAT scores and A Level points (a continuous variable) were also carried out.

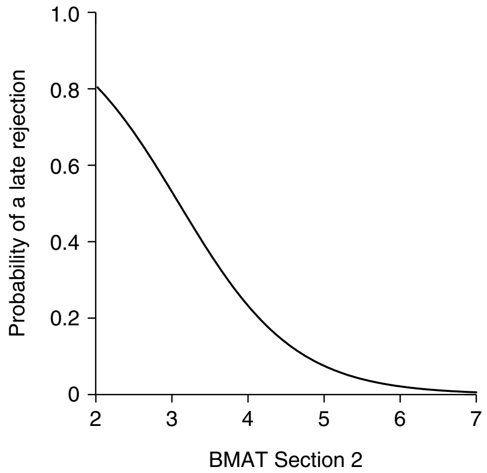
Correlations between BMAT scores and A Level points in the applicant group were 0.36 for Section 1, 0.36 for Section 2 and 0.26 for Section 3. All three BMAT sections also showed a strong positive relationship with the probability of achieving grades AAA in the applicant group (the probability being under 0.2 in applicants with Section 1 and 2 scores of around 3.0, compared to around 0.7–0.8, respectively, in those with Section 1 and 2 scores of approximately 6;

the probability was around 0.3 for a Section 3 score[6] of 4.5, compared to 0.6 for a score of 10.5). Importantly, BMAT Section 2 scores were a particularly strong predictor of failing to achieve at least grades BBB at A Level in the applicants who had been made an offer of a place conditional upon achieving these grades. It is particularly encouraging that all three sections were predictors of this outcome; Section 2's is perhaps unsurprising, given that applicants to biomedical school typically study two or more sciences at A Level.

Of the 178 candidates who had been made a conditional offer, 32 were rejected due to not achieving the BBB A level grade requirement. Those scoring around 3.0 on BMAT Section 2 had over a 0.5 probability of rejection at this late stage whereas those scoring around 5.0 had only a tenth of that probability (see Figure 6.9).

The use of BMAT scores as a potential early indicator of A Level performance is likely to become increasingly important given the proposed discontinuation of A Levels in their current form, which will increase universities' reliance on predicted A Level grades.

Figure 6.9 Logistic regression function showing the probability of a late rejection (failure to achieve A Level grades BBB) as a function of BMAT Section 2 scores (from Emery 2007c)



6.4 Chapter summary

In this chapter we outlined the importance of showing a relationship between test scores and other variables (criterion-related validity). We have detailed

[6] Emery (2007c) was conducted when Section 3 scores were awarded on a scale from 1 to 15.

the difficulties and limitations that are inherent to this field and outlined the approach adopted by Cambridge Assessment Admissions Testing with regard to measuring and reporting criterion-related validity.

For assessments used in selection such as BMAT, the relationship we are interested in primarily is with the future outcome that the test score is designed to predict (*predictive validity*). However, the relationship of BMAT with an outcome variable such as A Level performance may be considered predictive or concurrent depending on the intended use of BMAT scores in the selection process. We advocate a theoretical approach to the selection of outcome criteria and typically use measures of academic performance on biomedical courses (such as GPA) as criterion variables in our predictive validity studies. As predictive relationships are likely to be attenuated by range restriction, confounds on the outcome variables and the compensatory nature of the selection process, predictive validity is ideally measured during a pilot year for which BMAT scores are not considered in the selection process. However, where this is not possible we interpret uncorrected raw correlation coefficients according to recommended guidelines and take into account the selection criteria used by medical schools. This chapter also considered issues around the collection and collation of data, in particular the merits and limitations of multi-cohort and single-school studies. Predictive equity was discussed as an element of criterion-related validity that linked to consequential validity and test taker characteristics, which are covered in other chapters of this volume.

Finally, we described predictive validity work carried out on BMAT by Cambridge Assessment Admissions Testing. The studies presented in this chapter were conducted in collaboration with medical schools using BMAT as part of their admissions procedures. The results present good evidence of BMAT's predictive validity, demonstrating that scores on the test add value to biomedical admissions processes. Further work on the magnitude of some relationships between course components and specific test sections would add to this evidence, particularly for Section 3. The positive relationships identified so far are observable in single-site studies, despite the theoretical and methodological difficulties that attenuate observed correlations. Therefore test users can expect a degree of correlation between performance on BMAT and subsequent on-course performance.

Chapter 6 main points

- Tests used for selection are conceptualised as predictors, so predictive validity is more commonly investigated than concurrent validity in admissions testing.
- A range of issues weaken the relationships observed in predictive studies; so understanding the selection processes that were used can aid interpretation of results.
- The strength of predictive relationships between BMAT scores and course performance varies between institutions, courses and cohorts.
  - BMAT shows predictive validity across a range of courses and contexts, although the strength of correlations varies.
  - BMAT Sections 1 and 3 predict course outcomes in graduate entry into medicine.
  - BMAT predicts the likelihood of a student achieving their predicted A Level grades.
- Test users can expect positive correlations between BMAT scores and subsequent on-course performance, and also with likelihood to meet A Level offers.
- Future research may investigate concurrent validity in admissions tests, if suitable competency frameworks are developed.

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.