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

2 The biomedical school applicant: Considering the test taker in test development and research

Amy Devine
Research and Thought Leadership Group, Cambridge Assessment Admissions Testing
Lynda Taylor
Consultant, Cambridge Assessment Admissions Testing
Brenda Cross
University College London Medical School

2.1 Introduction

The test taker sits at the heart of any assessment event and ensuring that their needs are met is central to the fitness for purpose of an assessment. In this chapter we discuss the importance of the test provider having a sound understanding of the nature of the population for whom the test is intended. The BMAT test taker population is homogenous in some respects, because the majority of candidates are school leavers of a specific age range, ability level and language proficiency[1]. However, a substantial minority of those sitting the exam are referred to as 'non-traditional' applicants to medical school and test developers must be mindful not to disadvantage this subset of test takers. In other respects the candidature is more diverse. Several medical schools offer accelerated graduate-entry courses for applicants with an undergraduate degree in a scientific discipline. Also, increasing numbers of applications to medical school originate from outside the country where the medical school is based. Combined with growing use of BMAT in different locations, these factors mean that the educational backgrounds of applicants can be

1 The majority of BMAT candidates that apply to study undergraduate medicine at UK universities are home status students for whom English is their native language. Non-native English-speaking applicants to medical courses are typically required to demonstrate advanced English language proficiency (e.g. at the C1 to C2 level of the Common European Framework of Reference for Languages (CEFR), Council of Europe 2001).

quite different. An understanding of the test taker population informs meaningful evaluation of the different aspects of validity discussed in later chapters of this volume.

This chapter discusses how certain characteristics of the intended test taker population are taken into account in the overall design of the test. It also explains measures by which performance on the test is monitored and investigated to ensure fairness for different applicant groups, alongside and compared with outcomes from previous analyses.

2.2 The importance of test taker characteristics in assessment

The assessment literature often uses the term 'test taker characteristics' to describe a wide variety of features associated with the intended test taker population, which need to be taken into account when designing and administering a test. Test taker characteristics can include *physical* features (such as age and gender), *experiential* features (such as educational background or life experience) and *psychological* features (such as emotional state and motivation). Test designers need to have a clear understanding of the physical, experiential and affective features of the candidature for whom their test is intended. Test providers also need to have in place systems for investigating and monitoring test performance in relation to these factors since such features potentially influence testing outcomes. There are three main reasons for ensuring that a sound understanding and appropriate systems are in place.

First, for reasons of test validity and usefulness, it is essential that test content and format should be well matched to the intended test population (in this case applicants to medical, dental, veterinary and biomedical courses in higher education) and should be consistent with the intended purpose of the test and the scores generated. Where a test provider is informed about the nature of the target candidature for its test, and takes proper account of this in its test design, development and validation activity, the test is more likely to be fit for purpose. Combined with existing research knowledge about affective and psychological factors related to test performance, such as anxiety and risk-aversion, information about the population can support the design of constructs, tasks and scoring procedures.

Secondly, an awareness of test taker characteristics contributes significantly to test fairness. It is important to ensure that different applicant groups can access test content and formats without being unfairly disadvantaged due to demographic or background factors such as their age, gender, ethnicity or socio-economic group. In addition, any special requirements that may apply to individuals or subgroups within the intended test population, e.g. due to physical, psychological or emotional factors, need to be anticipated and addressed in an appropriate manner. Information on test

taker characteristics enables test providers to offer suitably modified tests (or testing accommodations) for those test takers who have temporary or permanent disabilities (e.g. a broken wrist or visual impairment). It also informs appropriate procedures to ensure fair treatment of those test takers who encounter some difficulty prior to or during the test which risks impairing their performance (e.g. bereavement, sudden illness, electricity failure).

Thirdly, systematic monitoring and analysis of test taker characteristics over time allows test providers to observe any changing trends within the test population and its characteristics. This information can inform future review and revision cycles of the test to ensure continuing validity and fitness for purpose.

The socio-cognitive approach outlined in Chapter 1 assigns a separate component to test taker characteristics within the overall test development and validation framework, thus maintaining a 'person-oriented' view of the testing and assessment process (rather than a purely instrument-focused view). At the same time, this focus on the test taker helps to ensure that the testing instrument meets the highest possible standards as far as matters of validity and fairness are concerned. The test taker characteristics component within the validation framework can be used to pose four specific questions (adapted from Weir 2005):

- What are the background characteristics of the test takers (age, gender, etc.)?
- Does the test make suitable accommodations for candidates with special needs?
- Are candidates sufficiently familiar with what they have to do in the test?
- Are candidates put at ease so that they are enabled to achieve their best?

These four questions are used by Cambridge Assessment to develop standard quality procedures and to design operational analyses for a test. The following part of this chapter (2.3) describes the standard quality procedures and operational analyses that relate to BMAT, and addresses each of these questions in turn. The four questions also frame targeted research studies which generate empirical evidence to confirm the validity and fairness of the test, examples of which are summarised later in this chapter.

2.3 BMAT and test taker characteristics

Collection of demographic data on BMAT candidates' characteristics

Key information on test taker characteristics is routinely collected for BMAT on multiple background variables and this information is matched to other variables in a variety of ways. The current BMAT registration process captures the following candidate background information for each test taker:

- test centre details (centre number, name, address and contact details)[2]
- candidate name (family and first names)
- gender
- date of birth
- Universities and Colleges Admissions Service (UCAS) ID number
- universities applied to with course code[3]
- requests for special needs access arrangements (where applicable, and with supporting justification).

Candidate background information is linked to BMAT test results (both test-level and individual item-level) via a unique BMAT candidate number allocated at registration. To support further research, additional candidate variables collected by UCAS during the university application process (such as ethnicity and socio-economic/participation of local areas (POLAR) group) can be matched to BMAT test results via candidates' UCAS ID number, also collected at registration. It should be noted that the data is typically more diffuse or sparse for graduate-entry medicine candidates than for the undergraduate population due to the intervening period between completion of school qualifications and sitting BMAT (for example, school information may not be available). These issues, combined with smaller sample sizes for graduate-entry cohorts, limit the analysis that can be conducted with graduate-entry BMAT applicants.

The BMAT registration form captures the candidate's signed consent that the data they provide may be used by Cambridge Assessment Admissions Testing and those institutions to which the test taker is applying, not just as part of the admissions procedures but also in associated follow-up research.

Routine analyses of BMAT performance by test taker groups

Shortly after each BMAT test session, results data for the whole cohort are analysed by gender, by school type and by UK/non-UK location. This provides a useful comparison for universities that use BMAT to understand the performance of their own cohort of applicants and how it might impact on admissions decisions.

Monitoring the composition of the BMAT candidature is another way of ensuring that the test remains correctly targeted and fit for purpose. A slightly higher proportion of female (approximately 56–59%) than male candidates have taken BMAT in every year. This reflects the distribution of gender

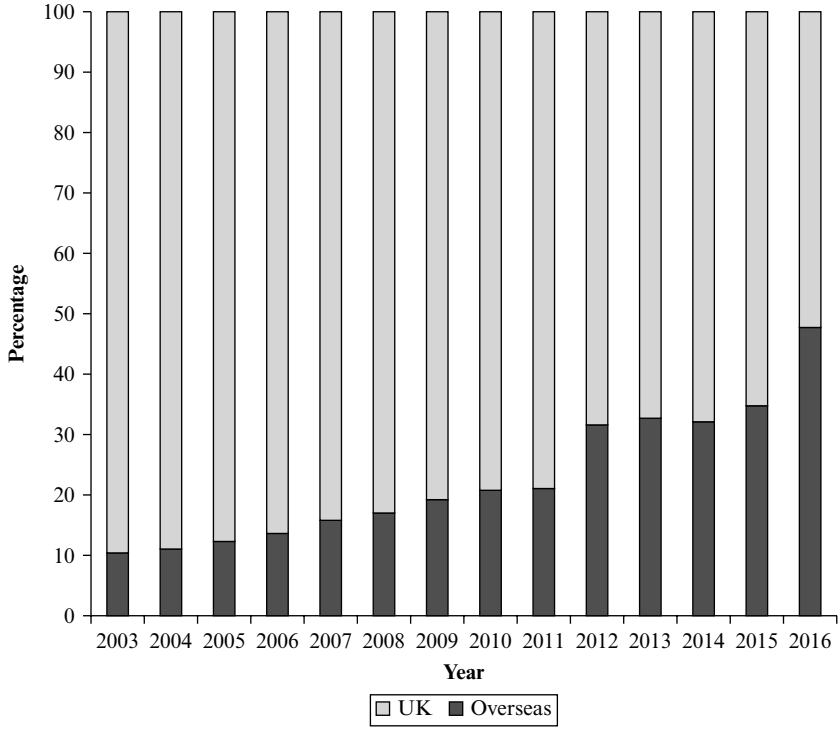
2 Because BMAT is normally taken in the test taker's own school or college, this information identifies the test taker's school type and location, thus permitting analyses of subpopulations according to these variables.

3 Restricted to universities and courses requiring BMAT.

amongst both those applying for entry to a medical course and the successful applicants, according to a report commissioned by the Royal College of Physicians (Elston 2009). The proportion of mature applicants has remained fairly stable over time, comprising approximately 10% of BMAT test takers. Monitoring the number of mature applicants and graduate-entry medicine applicants is important because these applicants may have additional needs which should be considered. For example, the limited number of BMAT test dates may necessitate mature candidates taking time from work or university study in order to sit the test; thus, Cambridge Assessment Admissions Testing may need to add further test dates or other testing arrangements if the proportion of mature candidates were to increase in the future.

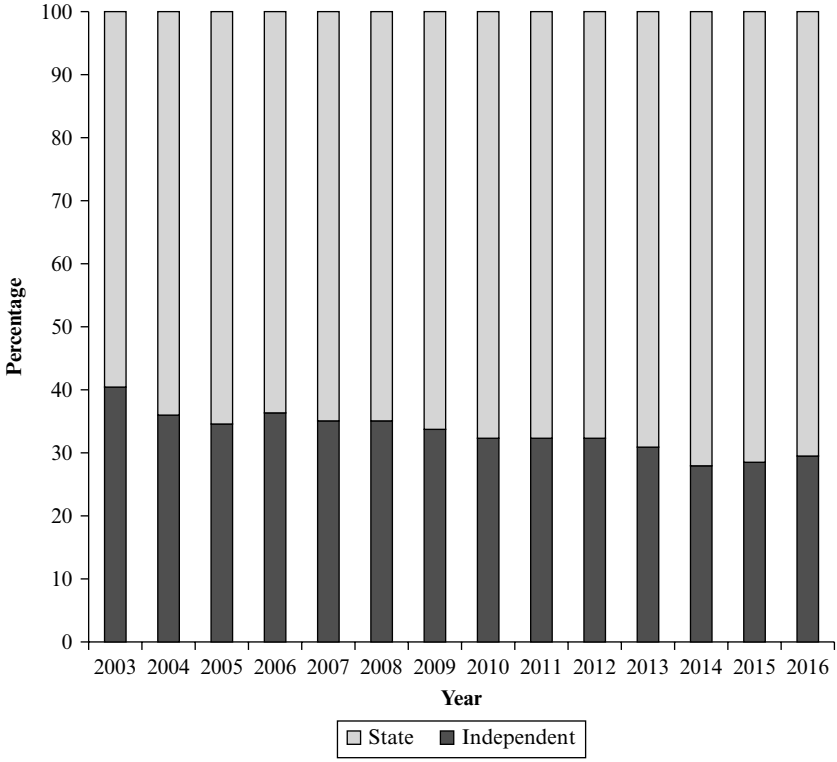
Figure 2.1 illustrates the composition of the BMAT candidature between 2003 and 2016 by location of test centre. The proportion of candidates from non-UK centres has increased from 10% to 48% over the 14-year period, showing a steady increase initially and then a steeper increase between 2011 and 2012 (coinciding with the first non-UK institution to use the test) and another increase between 2015 and 2016.

Figure 2.1 Centre location of BMAT candidates 2003 to 2016



UK school types are classified for analysis as belonging to either the state (government-funded) or the independent (private, fee-paying) sector. Those within the UK state sector are categorised into further subtypes, e.g. comprehensive, selective (grammar), sixth form college.[4]

Figure 2.2 School sector of BMAT candidates 2003 to 2016 (candidates from UK centres only)



There has been a gradual decrease in the proportion of independent (fee-paying) school candidates over the 14-year period, from around 40% of the UK-based candidates in 2003 to 29% in 2016 (see Figure 2.2). This possibly suggests an encouraging increase in the numbers of state school candidates accessing BMAT over time, or may reflect changes in the universities

4 The Academies Act passed in July 2010 made it possible for all maintained primary, secondary and specialist schools to apply to become academies. By 2016, 2,075 out of 3,381 secondary schools were academies, the number growing dramatically from 203 in May 2010 (www.bbc.co.uk/news/education-13274090). Progressive reclassification of comprehensive and grammar schools to academies should be borne in mind in analysis and interpretation of data collected for subtypes of state schools post 2010.

that use BMAT[5]. Nevertheless, as with applicants to medicine in general (Medical Schools Council 2014), independent school candidates remain over-represented in BMAT cohorts and state school candidates remain under-represented with respect to the pool of UK students attaining sufficiently high A Level grades for medical study (Emery 2010a)[6]. Changing this picture requires the ongoing commitment of the medical education community.

It is important to note that, to date, analyses of the school sector of BMAT candidates have understandably focused on the school type at which the candidate is engaged in, or has had their most recent educational experience. However, the school at which a candidate is studying or has completed A Levels or International Baccalaureate® (IB) is not necessarily the same as the one at which they completed General Certificates of Secondary Education (GCSEs) (or their equivalent). Trends in school applicants do not take into account the movement that takes place between the state and independent sector post-16 years of age. Whether a pupil continues their education at a given school beyond year 11 depends on a number of factors, including the existence of a sixth form, academic performance at GCSE (or the equivalent) and the subject options available to study at A Level/IB. It also depends on financial and other considerations.

Some high-achieving state school pupils, particularly those from less advantaged backgrounds, possibly identified by schools as 'gifted and talented' or through established links between the independent and state schools sector, move to independent schools for their sixth form education, often supported by scholarships and bursaries. Some others are assisted in their move to the independent sector by parents who perceive it as an investment, to improve their chances of achieving success at A Levels and beyond. The Independent Schools Council (ISC) recently reported that the number of pupils within its schools had reached its highest levels since 1974, with one in three receiving scholarships and bursaries (Independent Schools Council 2015). Barnaby Lennon, Chairman of the ISC, noted that 'one of the interesting features [of the current figures] is that it shows parents dipping into the independent sector for crucial stages of children's education' (Garner 2015).

5 Initiatives on behalf of universities, agreed as part of their Access Agreements with the Office for Fair Access (OFFA), to raise aspirations amongst under-represented groups, may have contributed to an increase in state school candidates accessing BMAT. It is hoped that the information and preparation materials made available to prospective candidates by Cambridge Assessment Admissions Testing, especially the free, online guides (discussed in this chapter and in Chapter 4) have increased confidence and provided reassurance to state school applicants.

6 The Department for Education (DfE) 2014 survey found that 79% of academies had changed or planned to change their curriculum and, of those that had, two thirds believed that the change had improved attainment. If this improved attainment extends to A Levels amongst the increasing number of academies, one might hope to see an increasing number of state school applicants achieving sufficiently high grades to meet medical school entry requirements in the future.

The biggest expansion in numbers was reported within the sixth form, with parents 'flocking' to the independent sector for private sixth form education, possibly as a result of increased affluence and/or worries about the state sector.

From 2015, as part of the ISC census, independent schools have been asked where their pupils were educated before joining their current school. The 2016 census showed that more than one quarter of pupils new to the independent sector came from state-funded establishments (Independent Schools Council 2016). Although movement into the independent sector was shown to occur at all ages, it was most pronounced at ages 16 and above, where 15% of pupils attended an independent school compared with 6% at age 11. The rise in numbers reported within the independent sector was not confined to UK students; it was also partly attributable to an increase in international student numbers.

Movement between state and independent sector, post-GCSE, is not unidirectional. There are a number of reasons why independent school pupils transition to the state school sector for their sixth form education. Reasons include: a greater range of A Level options, the desire for a change from boarding school or single-sex school, the draw of a greater social mix and broader life experience and the desire for a new start to revitalise interest in academic work, possibly after underperformance at GCSE. The cost of fees and parental concerns about perceived 'positive discrimination' in university admissions in favour of state school applicants, to meet government targets, are cited as important reasons for students leaving the independent sector post-GCSE. Whatever the reasons, it is clear to admissions tutors that a growing number of students are leaving independent schools after GCSEs and joining local state sixth forms.

A recent study, conducted for the ISC by the Centre for Evaluation and Monitoring at Durham University, concluded that independent school pupils performed better than state school pupils at GCSE (Ndaji, Little and Coe 2016). The average of the best eight GCSEs of independent and state school pupils differed by just under two GCSE grades before deprivation, prior academic ability and school-level factors were taken into consideration. The difference was reduced to 0.64 of a GCSE grade when these factors were controlled for, but the magnitude of the difference varied by GCSE subject. Nevertheless, the results suggest that attending an independent school is associated with the equivalent of two additional years of schooling by the age of 16.

In light of the movement between the state and independent sector post-GCSE, researchers and test developers, including Cambridge Assessment Admissions Testing, may wish to consider utilising school type at year 11 (data which is included on the UCAS application form) as well as school type at the time of application in future analysis of test taker characteristics.

In addition to monitoring the BMAT candidature by gender, school type and centre location, Cambridge Assessment Admissions Testing also monitors the proportion of candidates requesting special needs access arrangements for BMAT. Since 2003, between 1.2% and 2.9% of candidates in each cohort required extra time for special needs, and this type of provision typically accounted for the majority of access arrangements made. The proportion of candidates requesting extra time does not appear to be increasing over time.

BMAT data is also monitored for evidence of test bias by gender or school type. Item-level bias analyses are carried out annually for BMAT by both gender and school sector. A technique known as Differential Item Functioning (DIF) analysis (Holland and Thayer 1988) is used for this. DIF analysis compares the performance of two candidate subgroups (e.g. male and female) on individual test items, having matched the two subgroups on their overall test score as an indicator of ability. An item is flagged as potentially biased if one subgroup has a higher likelihood of getting that item correct than another subgroup when both are matched on overall test score. For a fuller discussion of DIF analysis, please refer to the key research study in Chapter 5.

Further analyses of BMAT performance by additional test taker variables (e.g. social deprivation indicators, candidates awarded extra time versus not) are carried out as larger-scale research projects. Examples of these studies are presented later in this chapter.

Information and preparation materials available to prospective candidates

An important factor affecting test performance is knowing what to expect on the day, so that candidates can concentrate on answering the questions rather than figuring out the test format. Candidates should familiarise themselves with BMAT prior to taking the test, and Cambridge Assessment Admissions Testing is committed to making preparation materials available free of charge on the BMAT website to obviate the need for candidates to pay for additional preparation. By providing clear, accessible and transparent information, the aim is to ensure that commercial test preparation does not offer additional insights to the information available on the BMAT website. Cambridge Assessment Admissions Testing staff also attend open days of selecting institutions to answer the questions of prospective BMAT candidates and their parents.

Test takers have free access to BMAT past papers and answers on the BMAT website, including worked explanations of answers to specimen questions and model responses to the BMAT Writing Task. The test papers allow candidates to become familiar with the clear test instructions given on the

front of each paper, and blank response sheets for all three test sections are provided to facilitate realistic practice.

BMAT is intended to be accessible for candidates without having to invest time learning large volumes of new material. Section 2 of BMAT ('Scientific Knowledge and Applications') is the only section that assumes subject-specific knowledge. However, it should be emphasised that BMAT Section 2 assumes core scientific knowledge in order to test the ability to *apply* that knowledge or principles to unfamiliar contexts, because this is what medical, dental and veterinary students have to do in their courses and beyond. In 2014, a review of BMAT Section 2 was undertaken to make more explicit the assumed science and mathematics knowledge, with the overarching aim of providing greater detail to test takers from a diverse range of educational and international backgrounds to support their preparation. In addition, revision guides have been created that prospective candidates can access, free of charge (*BMAT Section 2: Assumed Subject Knowledge guide*). The revision guides make clear the depth of knowledge required for each topic in a single electronic reference book (Chapter 4 of this volume gives a description of the process used to analyse core science curricula, update the BMAT specification and develop revision materials).

Published test preparation materials are also available for test takers. In 2010 (updated from the 2005 version) Cambridge Assessment, in collaboration with Heinemann, published a new set of test preparation materials entitled *Preparing for the BMAT: The Official Guide to the BioMedical Admissions Test* (Butterworth and Thwaites 2010). The book, which was authored and edited by specialists directly involved in the development and marking of BMAT, includes practice test questions and answers, together with guidance on approaching the test questions and worked examples. Its purpose is to familiarise test takers with the nature of the test, offering clear guidance about how responses will be scored so that candidates are given every opportunity to demonstrate the necessary knowledge and skills.

Understanding how candidates prepare for BMAT and the influence this has on their learning is an important consideration for the wider impact of the test. Candidates' use of the preparation materials provided by Cambridge Assessment Admissions Testing (and those from any other sources) has been the subject of BMAT research. This is described in Chapter 7. Cambridge Assessment Admissions Testing also carries out online surveys into test centres' and candidates' sources of information and preparation for our tests to better understand candidate needs, in order to guide the development of new support materials.

Access arrangements and special considerations

For test takers with special needs a range of access arrangements is available for BMAT, enabling test takers with disabilities to take the test on an equal footing as far as possible with other candidates:

- extra time (usually 25%)
- papers enlarged to A3
- supervised rest breaks
- other options on a case-by-case basis.

Access arrangements are requested in advance of the test by candidates' examinations officers (supporting evidence may be required). Where possible, Cambridge Assessment Admissions Testing adheres to the Joint Council for Qualifications (JCQ) recommendations for access arrangements and reasonable adjustments (see Joint Council for Qualifications 2016a) and BMAT candidates receive any arrangements that have been deemed necessary for their school examinations such as GCSEs.

There are also special considerations procedures in place to deal with unexpected problems that may arise immediately before or during the test, e.g. equipment failure, illness or accident on the day of the test, sudden interruption, excessive noise, etc. Requests for special consideration can be submitted by test centres on behalf of candidates within a fixed time period of the test date. An indication of the severity of the incident (as categorised by the Joint Council for Qualifications 2016b) experienced by the candidate is given to the receiving institution, so that they may take this into account, while maintaining any sensitive information about the candidate as confidential. No adjustments to candidates' marks are made by Cambridge Assessment Admissions Testing.

Psychological characteristics

One psychological factor associated with test performance is test anxiety, which is generally defined as fear and worry elicited by evaluative settings. Although there is a lack of appropriate normative data, research suggests that between 10% and 35% of school students and adults in post-secondary education are affected by test anxiety (McDonald 2001, Zeidner 1998). Moreover, females tend to report higher levels of test anxiety than males (Hembree 1988). Test anxiety is negatively correlated with test performance (Hembree 1988) and has been linked to lower performance in selection contexts (McCarthy and Goffin 2005). Item arrangement (specifically, whether test items increase or decrease in difficulty across a test) and time pressure have been associated with test anxiety and performance. Easy-to-difficult item sequences have been associated with lower levels of anxiety and better

performance than other test item sequences (Hambleton and Traub 1974), whereas increased time pressure is associated with lower performance, particularly in highly test-anxious students (Kellogg, Hopko and Ashcraft 1999, Plass and Hill 1986).

Item arrangement and time pressure are considered in BMAT test construction. For example, as far as possible, BMAT items are ordered to increase in difficulty over each test section[7], in order to minimise anxiety at the outset of the test. Moreover, the number of items and the number of complex or time-consuming items in BMAT was adjusted in its first years of administration to ensure the timing of the test is sufficiently challenging but not unnecessarily stress inducing (see Chapter 4).

In addition, as mentioned above, Cambridge Assessment Admissions Testing offers BMAT preparation materials online, such as the test specification, BMAT past papers and answers. This provision potentially reduces test anxiety by enabling candidates to familiarise themselves with the test format and undertake realistic practice prior to sitting BMAT (see Chapter 7 for further discussion of BMAT candidates' use of preparation materials).

Another psychological factor which must be considered is risk aversion. There is evidence to suggest that males and females differ in the extent to which they are willing to take risks in high-stakes tests; several studies have shown that females are more likely than males to omit responses to item types in which incorrect responses are penalised (Baldiga 2014, Hirschfeld, Moore and Brown 1995, Kelly and Dennick 2009). However, there is not consistent evidence of gender bias in multiple-choice questions (MCQs) which do not employ this scoring method, particularly when it has been investigated with large-scale studies (Arthur and Everaert 2012, Bramley, Vidal Rodeiro and Vitello 2015, Buck, Kostin and Morgan 2002, Du Plessis and Du Plessis 2009). Collectively, the results suggest that negative marking may lead to gender bias in multiple-choice tests; therefore, this score-awarding method is not employed in BMAT scoring (see Chapter 5 for further details of BMAT scoring).

Small but significant group differences (including gender differences) in BMAT scores have been found (see section 2.4). However, it should be noted that group differences are fairly ubiquitous in medical admissions testing. For example, males have been found to outperform females and native English-speaking candidates have been found to outperform non-native English-speaking candidates on the United Kingdom Clinical Aptitude Test (UKCAT) (Tiffin, McLachlan, Webster and Nicholson 2014). Performance differences by gender, race/ethnicity, or socio-economic status have also been reported for the Medical College Admission Test (Association of

7 Note that items for each subtype are interspersed in each test section, thus there is some variation in the difficulty of items across subtypes but overall, items tend to increase in difficulty across the section.

American Medical Colleges 2016), the Erasmus MC Medical School cognitive tests (Stegers-Jager, Steyerberg, Lucieer and Themmen 2015), and the Undergraduate Medicine and Health Sciences Admission Test (Griffin and Hu 2015), used by medical schools in the US, the Netherlands and Australia respectively. Whilst the method effect (i.e., format of examination questions) is an important factor to consider when investigating the sources of these group differences, there are many other factors which may contribute to group differences in performance on medical admissions tests.

BMAT candidates are a self-selected population and tend to represent the highest performing students across a range of subjects relevant for medical study. Gender differences in science, mathematics and reading performance are more pronounced at the upper end of the performance distribution (Hyde, Lindberg, Linn, Ellis and Williams 2008, Nowell and Hedges 1998, Stoet and Geary 2013). Moreover, there are many psychological, social and cultural influences on school subject choice and performance which may contribute to self-selection into medical study (Eccles, Adler, Futterman, Goff, Kaczala, Meece and Midgley 1983, Eccles 2011). Thus, group differences in performance on BMAT may reflect factors outside of the test. Nonetheless, Cambridge Assessment Admissions Testing periodically monitors BMAT for item-level bias (see the section on item-level bias analyses in this chapter, also described in more detail in Chapter 5).

2.4 Research on test taker characteristics

Key study – Investigating the predictive equity of BMAT (Emery, Bell and Vidal Rodeiro 2011)

A key piece of research into the fairness of BMAT for selecting different test taker groups was published in 2011 by Emery, Bell and Vidal Rodeiro. This investigated the relationships between medicine applicants' background characteristics (gender, school type, neighbourhood deprivation etc.) and the following: their BMAT scores, whether they were offered a place of study or rejected, and, for those admitted, performance on their first year course examinations.

Test fairness does not require equal group performance (*Standards*, 2014). However, psychometric definitions of test bias rely on the central notion that different groups of candidates *with the same standing on the construct of interest* should attain, on average, the same test score. Group differences in test scores that reflect group differences on the construct of interest are not problematic but those that exist due to irrelevant sources of variance are.

When test scores are used to predict a future outcome, as in the case of BMAT, then scores (technically the *use of* scores) can be regarded as biased against a particular group if they *under-predict* future performance for that

group (*Standards*, 2014). That is, the score implies a lower level of ability than is really the case. Scores can be regarded as biased in favour of a particular group if they *over-predict* future performance for that group (that is, the score implies a higher level of ability than is really the case). This is known as *predictive bias* (Cleary 1968) and definitions of bias or a lack of bias in the admissions testing context generally rely upon the analysis of this.

Testing for predictive bias involves using regression analysis where the criterion measure (course outcome) is regressed on the predictor variable (admissions test score), subgroup membership and an interaction term between the two. If a particular admissions test score for two groups of candidates reflects the same underlying ability on the construct of interest (i.e. potential for success on the course) then we would expect predicted course performance to be the same between them, other things being equal (Cleary 1968). Differences in the regression slope and/or intercept between different test taker groups indicate predictive bias. In Emery et al (2011), therefore, the fairness of BMAT for student selection was investigated by determining whether a particular set of BMAT scores predict the same future course performance, on average, for different groups of test takers.

Three successive years of undergraduate medicine applications data to the University of Cambridge were used for the analyses. Mature and non-UK applicants were excluded from the study so that the same admissions criteria could be assumed to have applied to all those included. Test taker characteristics included in the study were gender, school type (comprehensive versus each of the following: independent; grammar (selective); sixth form/tertiary colleges; FE colleges) plus a range of social (neighbourhood) deprivation indicators. Neighbourhood deprivation indicators were downloaded from the Office for National Statistics (ONS) website and were matched to candidates' school postcode information (home postcode information was not available). Measures included income, employment and education deprivation indicators.

Results showed that, despite some differences in applicants' BMAT performance by background characteristics (e.g. by school type and gender), BMAT scores predicted average first year examination marks equitably for all the background variables considered. Regarding performance differences, the male applicants in these three combined Cambridge cohorts scored higher than the female applicants on BMAT Section 1 (0.19 of a BMAT point) and on Section 2 (0.23 of a BMAT point). Section 3 scores were not included in this analysis because the University of Cambridge did not use Section 3 scores in selection in these test years (2003–05), instead considering candidates' responses as a qualitative piece of evidence. The largest difference relating to BMAT scores in these cohorts was for comprehensive versus independent school applicants on BMAT Section 2, with the latter group scoring 0.34 of a BMAT point higher, on average, than the former.

Associations between BMAT scores and the neighbourhood deprivation variables were weaker or non-significant, with the largest effect found for one of the neighbourhood employment indicators (here, each 1% increase in adults on the lowest social grade in the neighbourhood was associated with only a 0.02 BMAT point decrease on Section 2).

However, and crucially for BMAT, the relationship between BMAT scores and future course performance (year 1 examination average percentage mark) did not differ for any of the test taker groups or by any of the continuous background variables. Despite differences in BMAT scores between groups, a given set of BMAT scores predicted the same medicine course examination result, on average, for all test takers regardless of group. This provides important evidence that BMAT scores mean the same for different test taker groups. That is, the empirical evidence suggests that candidates with the same BMAT scores have the same standing on the construct of interest regardless of their gender, school type or level of social deprivation.

In conclusion, differential performance on a test by different candidate groups, even if taken to be truly representative, is not a legitimate way to measure test bias (*Standards*, 2014). The real issue is whether the score differences between test taker groups reflect genuine differences between them on the construct of interest (as the analysis suggests here) or are a result of construct-irrelevant sources of variance that result in systematically higher or lower scores for certain groups. A given test score should reflect a certain level of ability regardless of group membership. For admissions tests (or any measure used in selection), scores should predict future performance equitably provided that other factors such as motivation are equal between test taker groups. Unlike bias, however, *fairness* is not a psychometric concept and views about the fairness of admissions procedures will vary even given unbiased measures. The equitable treatment of all applicants, however, is key to most definitions.

An overview of other research

In light of the importance of monitoring BMAT for fairness on an ongoing basis, a range of research studies have been carried out into both the performance of different test taker groups and the provision of suitable arrangements for candidates with disabilities.

Item-level bias analyses (Emery and Khalid 2013a)

A key research study into item bias in BMAT is discussed in detail in Chapter 5. However, to summarise the findings here, no evidence of DIF was found for any BMAT item by either gender or by school sector in the three consecutive test cohorts examined. This suggests that there is no evidence of bias in

any BMAT items and therefore they do not advantage, for example, males over females or private school candidates over state school.

BMAT test taker characteristics and the performance of different groups 2003–12 (Emery 2013b)

Candidates classed as mature applicants scored lower than non-mature applicants on BMAT Section 1 in most test years (small effect sizes), on Section 2 in all test years (medium effect sizes) and on Section 3 in four years only (small effect sizes). Differences in scores between groups of applicants can be a concern to selecting institutions (particularly those that use the test as a hurdle to the interview stage) due to the possible impact on the composition of those admitted. Evidence from this research study clarifies that small but statistically significant differences in BMAT performance have been found by both gender and by school sector, with male candidates and those from independent schools performing slightly higher on the two MCQ sections of the test. Conversely, female candidates tended to perform slightly higher than male candidates on Section 3. Effect sizes often appear to be larger for Section 2 of the test (although only small to medium).

Differences in test performance do not, in themselves, equate to test bias: they may reflect genuine differences on the construct of interest between different groups of applicants. The latter has been investigated using regression techniques and the results suggest that candidate-group differences in BMAT performance reflect genuine differences in how they are likely to perform for the course of study. Additionally, DIF analyses of sets of test items can clarify whether gender and other differences are the result of item bias and whether they should therefore be regarded as a genuine cause for concern by test users (please refer to Chapter 5).

The performance of mature candidates on BMAT Section 2 is of interest to institutions using BMAT, given the longer time interval since their GCSE studies (or equivalent) at school or college. Emery (2013b) confirmed slightly lower scores for mature applicants on both Sections 1 and 2 in most years. Recent analyses comparing the BMAT performance of *graduate* applicants to the under-21s has not replicated this difference in Section 2 scores, suggesting that time out of education may be the causal factor.

Investigating BMAT for candidates with disabilities (Ramsay 2005)

A research project funded by Higher Education Funding Council for England (HEFCE) was carried out by the University of Cambridge Disability Resource Centre (Ramsay 2005). This mixed methods study looked at various admissions assessments introduced into the undergraduate admissions process by the University of Cambridge, including BMAT, and whether these appeared to disadvantage students with disabilities.

Secondary quantitative analysis of BMAT test data (originally collected

by Cambridge Assessment Admissions Testing) showed that candidates with disabilities did not appear to be disadvantaged by the tests: the marks of candidates who requested access arrangements were not uniformly lower than those of other candidates, nor was there an imbalance in their success rate in being offered a place of study. Qualitative methods were used to investigate issues beyond test performance, such as the information provided to disabled candidates, the test registration process, responsibility for ensuring that access arrangements are put into place, travel to test centres, etc. Finally, a mock test of thinking skills items akin to those in BMAT Section 1 (all MCQ) was taken by a small group of admitted students with a range of disabilities, with the usual access arrangements put in place, including extra time. Participants were interviewed about their experiences with the mock test (and any actual admissions tests they had taken), such as any questions they found particularly difficult and any issues with the test format or content. Both the participants and admissions tutors were interviewed about their thoughts on whether admissions tests would aid in student selection.

Results gave no cause for concern in the access-arrangement group in terms of mock test performance, reported issues with the mock test or views of fairness regarding the introduction of admissions tests for student selection. Interview feedback was positive, with comments from the access-arrangement group typically stating that their disability had not been a problem for the (modified) test, or explaining why the extra time had been necessary for them. However, one interviewee commented that the BMAT Writing Task may be much more difficult than MCQ items for a candidate with dyslexia: 'The main part of my disability is expressing things . . . (in the MCQ format) . . . it is expressed for you.' Positive views on the utility and fairness of thinking skills tests for student selection were received from both the mock test participants and the admissions tutors. However, the author made a number of recommendations regarding issues that were 'broader than the test paper', such as the provision of information on applying for access arrangements and the accessibility of test centres for candidates with physical disabilities. The report concludes by emphasising that qualitative research into the experiences of test takers with special needs can highlight how individual the difficulties resulting from disability can be, and how it is hoped that understanding of the disability issues relevant to assessment continues to grow.

2.5 Chapter summary

In this chapter we have discussed how test taker characteristics are taken into consideration in the overall design of the test. The key study illustrated the importance of investigating the predictive equity of a test, in order to monitor any potential bias. Routine analyses of BMAT performance show

that different candidate groups do not necessarily perform equally on the test, but research evidence shows that BMAT predicts course performance equitably for different test taker groups. It is also important to monitor the background characteristics of the population for whom a test is intended. Due to recent trends for movement of pupils between state and independent sectors post-GCSE, future analysis of test taker characteristics should consider utilising school type at year 11 as well as at the time of application. The procedures and research carried out on BMAT aim to ensure that the test is as fair as possible for different candidate groups, including those with special needs. This is vital given the high-stakes nature of BMAT. The test information and wealth of free preparation materials provided to BMAT candidates by Cambridge Assessment Admissions Testing aim to level the playing field for those from different backgrounds and allow all test takers to perform to the best of their ability.

Chapter 2 main points

- Monitoring the demographics of test takers can inform test development and revision.
- Information about the test taker population supports investigating various aspects of validity.
- Differences in performance between groups do not necessarily indicate bias.
- Care must be taken to understand the contexts and categorisations of different groups for a test.

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.