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

7 The consequences of biomedical admissions testing on individuals, institutions and society

Sarah McElwee
Research and Thought Leadership Group, Cambridge Assessment Admissions Testing
Molly Fyfe
Research and Thought Leadership Group, Cambridge Assessment Admissions Testing
Karen Grant
Lancaster Medical School

7.1 Introduction

This chapter explores consequential validity, which refers to the impact that a high-stakes test, such as BMAT, has on all its varied stakeholders (including candidates, teachers and universities), on teaching and learning, and on society more broadly. Weir's (2005) socio-cognitive framework considers the social consequences of test use as part of overall validity and Cambridge Assessment Admissions Testing also adopts this position, treating the consequences of using an admissions test as part of overall validity (Messick 1995).

Box 7.1 Definition of consequence in educational assessment

Consequences: The outcomes, intended and unintended, of using tests in particular ways in certain contexts and with certain populations.

(*Standards* 2014:217)

In this chapter we describe the way that Cambridge Assessment Admissions Testing investigates the social consequences of BMAT and supports positive impact (*impact by design*). The features of BMAT that support student revision and promote valuable thinking skills ('positive washback')

are discussed. This includes a description of how stakeholder needs are addressed through collaboration with institutions using BMAT. At a time of heightened media scrutiny of fair access to higher education, the role of BMAT in supporting transparent admissions processes to heavily over-subscribed courses is outlined by Professor David Vaux, an Admissions Tutor at the University of Oxford. Two key studies are presented in this chapter. The first study details findings from a survey of BMAT candidates on their test preparation activities, which was conducted to understand how preparing for tests like BMAT can impact upon student learning and test performance. The second study explores candidates' attitudes towards admissions tests and the wider process of applying to study medicine, again using survey methods.

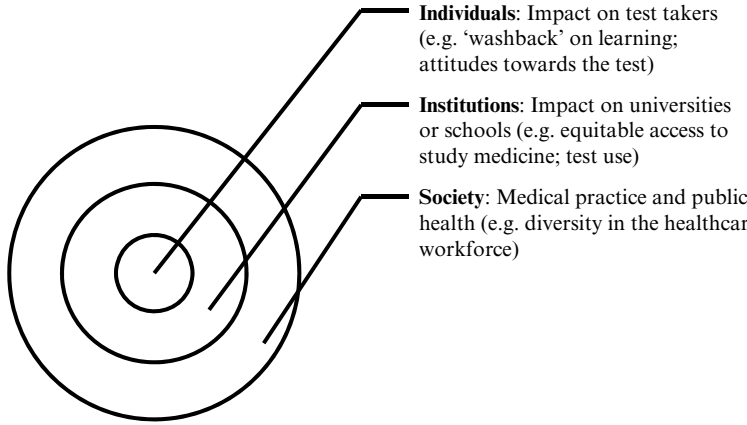
7.2 Consequential validity in medical selection

Within the field of medical education, consequential validity tends to be viewed as issues relating to the interpretation and use of test scores (Downing 2003). As will be fully described in the following section, we adopt a broader view on consequential validity that not only includes score use and interpretation, but extends to important issues such as test preparation behaviours, equity and stakeholder perceptions. We feel that this approach is particularly important when considering consequential validity within the context of admissions to medical study.

High-stakes testing for university admission directly affects the choices, careers and experiences of thousands of young people aiming to follow a particular educational path. The institutions that use these tests are also affected; at a micro level (in the effect it has on their admissions decisions and the performance of the cohort they select) and in a wider sense (in linking their reputation to the assessment). More broadly still, at the societal level, issues of social justice, fair access and public confidence in assessment are all relevant to high-stakes testing, and in particular to admission to medical school.

The social impact of BMAT extends to issues such as the diversity of the physician workforce and public health. The British Medical Association (BMA) argued in 2009 that 'doctors should be as representative as possible of the society they serve in order to provide the best possible care to the UK population' (British Medical Association 2009). The General Medical Council (GMC) reported in 2011 that the medical profession has made significant strides in terms of diversity and change in recent years, with large increases in the number of doctors who are female and from ethnic minorities. However, in 2012, Higher Education Statistics Agency (HESA) data demonstrated that the proportion of applicants from lower socio-economic groups gaining access to medical study was still lower than desired (Milburn 2012). Under-representation of physicians from lower socio-economic

Figure 7.1 The context of BMAT scores



backgrounds in the workforce has a profound impact on society as these doctors are those most likely to work with underserved patient populations (Dowell, Norbury, Steven and Guthrie 2015).

The processes of selection to medicine are complex, with many medical schools using a wide range of evidence, including school academic performance, work experience, 'traditional' (panel) interviews, multiple mini-interviews (MMIs) and teacher recommendations. As a key part of this process, it is important that admissions tests such as BMAT do not act as a deterrent to application, particularly in regard to the entry into medicine of students from lower socio-economic backgrounds.

Defining consequential validity

Consequential validity is conceptually distinct, though related to, the other types of validity discussed in this book. Issues such as cognitive validity, scoring validity and context validity relate primarily to the quality of a test as a measurement instrument ('technical quality') and are the responsibility of the test developers to address (Newton and Shaw 2014). In contrast, consequential validity is concerned with the impact that a test has on an individual, institutions or society ('social value'). Consequential validity must attend to socio-cultural contexts and policies relating to test use. Stakeholders, such as university departments, largely determine how the tests will be used in practice, and so influence the consequential validity of BMAT. Consequential validity is also influenced by the test design, schedule and preparation practices. For BMAT, the approach adopted by Cambridge Assessment Admissions Testing influences the consequential validity of the test, because decisions made by the test developer can impact how the test is used.

Approaches to validation frequently draw on frameworks or models to operationalise validation processes. Weir's (2005) socio-cognitive model, used throughout this volume to frame the validation evidence for BMAT, includes consequential validity as a crucial piece of evidence for scrutinising the fitness for purpose of a test. This aspect of Weir's model is influenced by Messick's (1989) concern with the 'consequences of test use'. Messick argued that any model that did not account for consequential validity was inadequate, as it failed to account for 'both evidence of the value implications of score meaning as a basis for action and the social consequences of score use' (Messick 1995:741).

Box 7.2 Messick's definition of validation

Validation is empirical evaluation of the *meaning* and *consequences* of measurement.

(Messick 1995:742, emphasis added)

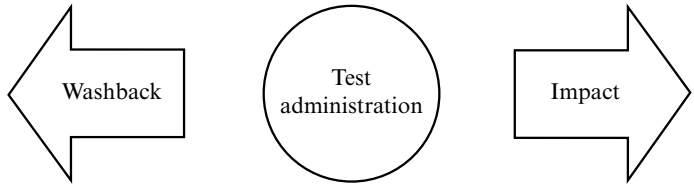
While there is consensus that the social consequences of test use are crucial to consider, there is debate over whether these should be included in a 'unified' validity framework (as Weir proposes) or whether 'technical quality' and 'social value' should be conceptualised as separate issues. In the *International Handbook of Research in Medical Education*, Shea and Fortna (2002:110) summarise this issue by stating that 'no-one disagrees that the social consequences of test uses (and misuses) are important. The dispute is whether to call it "validity" or not'. In the wider educational assessment community, Cizek (2012) has argued that ethical and social considerations, such as those discussed in the present chapter, do not fall in the realm of validity. Others have narrowly defined validity to specifically exclude ethical and social evaluations regarding how test scores are used (Borsboom, Mellenbergh and van Heerden 2004). However, even these critics of consequential validity concede that the suitability of a testing procedure depends on more than the properties of the test itself. The *Standards* also recognised that the consequences of introducing an assessment are important to consider when evaluating a test (see Box 7.1).

For a more in-depth discussion of validity theory and the cases made for and against consequential validity, the reader is referred to Newton and Shaw (2014), who treat this topic in some detail. In the present chapter, we adopt the approach advocated in the socio-cognitive framework, by classing these issues as part of validity that need to be evaluated. Like Weir (2005), we advocate treating consequential validity as equal in status to other aspects of validity that are systematically and regularly considered. Consequential

validity must be considered alongside other measures of test quality to ensure 'fitness for purpose' as it is possible to have a test that is an accurate measurement instrument, but that has negative impacts due to how it is used (Cronbach 1988). Included within this conceptualisation of consequential validity is an admissions test's impact on the behaviours of potential applicants and on universities using the test.

Consequential validity encompasses three elements: *washback*, *impact* and *differential validity* (Figure 7.2). Washback is effects that the test has on potential test takers or institutions before it is administered, for example, through preparation behaviours. Impact of the test occurs after it has been administered, for example through how test scores are used in the admissions process. In the admissions testing context, because consequences arising from use of a test often impact on future admissions cycles and test administrations, washback and impact can interrelate. For example the perceptions of people who take the test about its fairness may go on to influence how future generations of test takers will view the exam. In particular, views towards an admissions test, and whether these may influence a prospective student's decision to apply to a course, are important aspects of consequential validity.

Figure 7.2 The directionality of impact and washback



Washback

Washback refers to the influence that an examination has on educational practices. The adage that 'assessment drives learning' is well established in medical education (Newble 2016); 'washback' is a term used widely in the literature on language testing to describe this phenomenon. There is evidence that tests shape learners' preparation behaviours, educational materials, the teaching they receive and the curriculum they follow (Green 2007, Luxia 2007, Newble and Jaeger 1983, Saville and Hawkey 2004). Washback can be positive when test preparation encourages the acquisition of knowledge and skills which are beneficial beyond the context of the test. Conversely, negative washback refers to study behaviour that focuses only on 'learning the test'. Examples of negative washback occur when a test directs students to concentrate on narrow aspects of the curriculum, rewards attempting to 'question spot', or encourages focus on test-taking strategies at the expense of learning. Indeed, recent A Level reforms in

England were introduced to combat such negative washback effects that were perceived to be adversely impacting on learning and understanding: in April 2013, David Laws, the Schools Minister, stated: 'They [school students] and their teachers have spent too much time thinking about exams and re-sitting them, encouraging in some cases a "learn and forget" approach' (Long 2017). To maximise positive washback for candidates it is important to emphasise the relevance, importance and attainability of items in the test and to ensure they are appropriate for the test takers (Green 2003, 2006, Hughes 2003).

Hughes (2003) suggests that positive washback in high-stakes tests can be achieved by testing the abilities whose development you want to encourage, by sampling widely from the curriculum and by ensuring that the test is known and understood by students and their teachers. Green (2003) adds some further details that contribute to positive washback: that success on the test should be perceived to be both important and difficult (but attainable), and that these perceptions are shared by other test takers.

It is worth noting that washback from BMAT will occur in different ways than would be expected in a language testing context, or the context of other high-stakes exams, such as General Certificates of Secondary Education (GCSEs). With BMAT, there is no expectation that schools provide specific preparation for the test, and in fact a key concern of the universities using BMAT is that preparation should not entail significant new learning. However, one could argue that preparing for Section 2 of BMAT would encourage candidates to revise GCSE maths and science, and learn how to apply this knowledge in unfamiliar contexts, enhancing their pre-existing knowledge and developing skills that will be useful for their further study (A Levels and beyond). Furthermore, only a relatively small percentage of students will sit BMAT. Thus, it is unlikely that BMAT will influence the wider system of secondary education, and washback will be observed in terms of impacts on students' out-of-school activities (such as self-directed test preparation) and learning.

Impact

Impact, as described earlier, is the effect that the test has on the full range of stakeholders, and on society more generally. Test takers and selecting institutions are those affected most directly, as the results influence decisions about their future study paths and careers, and their academic cohorts, respectively. Additionally, schools, parents, and national medical and veterinary associations represent just some of the other groups impacted by tests such as BMAT in the wider social sphere.

Perhaps one of the most important impacts of admissions tests for medical study is the observed effect that selection might have on student learning and achievement. Kreiter and Axelson (2013) note that:

> Although effective educational interventions typically produce only small gains in learning, usually with effect sizes of .20 or less, evidence-based selection is comparatively far more powerful. In fact, when well designed, selection procedures in medical education can achieve performance gains easily exceeding 1 standard deviation (Kreiter and Axelson 2013:S51).

Entry to medical school is highly competitive: for 2017 entry, Universities and Colleges Admissions Service (UCAS) received 19,210 applications for approximately 6,000 medical school places (Universities and Colleges Admissions Service 2016). Consequently, medical schools aim to select those who are best suited to studying medicine, and have the best chance of successfully completing the medical degree programme. The role that an admissions test can play in selecting medical students and the positive impact that it can have on student learning and achievement is illustrated in the study by Reibnegger, Caluba, Ithaler, Manhal, Neges and Smolle (2010). Comparison of cohorts of medical students before and after the introduction of an admissions test into the selection process found that the probability of success at medical school was dramatically increased when students were selected using an admissions test compared to those admitted under an 'open' system. The reasons for this increased success rate were not explored but could include: students who performed better on the admissions test were better suited to the intellectual challenge of studying medicine; students who performed better in the test were more motivated to become doctors, and had invested more time and effort in preparing for the admissions test (see Wouters et al 2016). Whatever the underlying reason for the effect, it is evident that selecting a student body with a higher probability of educational success will have a positive impact on the medical school, as well as on individual students.

The impact of an admissions test is only partly explained by the test itself; it will also be determined by the policies surrounding its implementation and the way in which the test scores are used to select candidates. The study by Reibnegger et al (2010) illustrates how a change in government policy can have an impact on both the medical school and its students. Any impact on educational success and dropout rates will also be influenced by the way in which institutions use admissions test scores in their selection processes, and this varies between institutions (see the next section). Therefore stakeholders, including universities and regulatory bodies, play a key role in shaping test impact through the decisions they enact around test use.

Differential validity

Weir (2005) also includes *differential validity* as an aspect of consequential validity, relating to factors that differentially affect the performance of different groups of candidates. Issues of differential validity may pertain to difference in test-related behaviours, attitudes or outcomes by gender, ethnic,

socio-economic or other demographic groupings. Although there is diversity in the selection methods employed at different UK medical schools, they universally include academic achievement, in terms of GCSE and A Level grades. This may not be surprising as there is evidence that past academic achievement is a useful predictor of success at medical school (Patterson et al 2016). However, school academic achievement is influenced by factors unrelated to potential: those from lower socio-economic groups tend to underperform relative to their more affluent peers (Blandon and Gregg 2004) even though this difference disappears once they enter higher education (Hoare and Johnston 2011).

One of the rationales for using an admissions test is to provide a standardised measure that levels the socio-economic and educational inequalities inherent in a pool of applicants. It is therefore crucial to evaluate admissions tests for differential validity to ensure that the tests do not reinforce inequity.

Bias is a key issue to consider in differential validity and is defined as score differences between groups that are not related to the construct being assessed. There are different ways of investigating bias, such as Differential Item Functioning (DIF) (Chapter 5) and predictive equity (Chapter 2). In the case of BMAT, there are persistent differences in mean scores, with males performing slightly better than females in Sections 1 and 2, and those from independent schools performing slightly better than those from comprehensive schools overall (see the key study in Chapter 2). However, there is no evidence of these issues being due to test bias, and the conclusion is rather that they are due to construct-relevant variance between the groups, which likely arises from a larger spectrum of socio-cultural influences which impact on students throughout their lives.

While the observed differences in BMAT scores are not attributable to test bias, Cambridge Assessment Admissions Testing is conducting further research to investigate these issues. In the two key studies presented in this chapter, we investigate issues of perceived fairness and test preparation, looking beyond statistical understandings of bias to other issues, which may affect the differential validity of an assessment. Findings from this research have helped shape our approach to supporting test preparation that is equitable, as will be described next.

Cambridge Assessment approaches to consequential validity

Impact by design

Consequential validity is considered in the design of BMAT test materials using the principle of *impact by design* (Saville 2012). This aligns the practice of Cambridge Assessment Admissions Testing with the Cambridge English approach to designing language tests. According to this position, test design

and production processes should consider right from the outset the potential uses of the test, in order to maximise positive test impact for candidates. Test developers should also anticipate and mitigate negative impact as far as possible. By following four maxims (see Table 7.1) the positive impact of the test is enhanced as far as possible.

Table 7.1 Impact by design

Maxim 1	**PLAN** Adopt a rational and explicit approach to test development
Maxim 2	**SUPPORT** Support stakeholders in the testing process
Maxim 3	**COMMUNICATE** Provide comprehensive, useful and transparent information
Maxim 4	**MONITOR and EVALUATE** Collect all relevant data and analyse as required

By definition, impact by design principles are integrated throughout the test development and validation cycle (see Chapter 1 for the phases of the cycle), and how the test will be used by stakeholders is considered early in the planning phase. The substantial role that early users of BMAT had in defining the test meant that the intended uses were explicitly included in initial plans and subsequent reviews of BMAT. Cambridge Assessment adopts Saville's (2012) maxims by aiming to support and communicate with stakeholders continuously. The processes for this are outlined in the following portion of the chapter, alongside research that monitors and evaluates the impact of the test.

The socio-cognitive framework

Cambridge Assessment Admissions Testing conducts studies that evaluate the consequential validity of admissions tests. Research in this area tends to be naturalistic, that is, it is focused on exploring existing practices and perceptions. Research and practice on BMAT's consequential validity is framed according to the socio-cognitive validation framework proposed by Weir (2005), which can be used to pose five guiding questions as presented in Box 7.3.

Box 7.3 Questions on consequential validity (Weir 2005)

1. Are actions based on test scores appropriate?
2. Is there a washback effect in the classroom (positive or negative)?
3. Is there any evidence of differential validity?
4. How are candidates preparing for the test?
5. How is the test perceived by stakeholders?

The rest of this chapter considers issues relating to consequential validity, using the criteria proposed by Weir (Box 7.3) as a framework. Thus, we describe the use of BMAT scores by universities, consider washback in the context of BMAT, and present research investigating applicants' preparation behaviours, and their perceptions of the test.

Evidence of differential validity was addressed in Chapter 5, in which work on investigating and preventing bias in BMAT was presented. As part of the work presented here, we discuss research studies that investigated how differential access to test preparation material impacts on test performance, and also how students from different backgrounds perceive BMAT.

7.3 Are actions based on test scores appropriate?

Appropriate score use centres on two issues: the reliability of the test as a decision-making instrument, and how scores are used in practice to make selection decisions.

Based on research presented in earlier chapters of this volume, we know that BMAT scores can effectively be used to support admissions decisions. There is a close relationship between the score interpretation aspect of consequential validity and the criterion-related aspect of validity, which is discussed in more detail in Chapter 6. Ensuring that test scores mean the same thing for all test taker groups (as discussed in Chapter 2, on test taker characteristics, and in Chapter 5, on scoring validity) is another facet of consequential validity. Chapter 2 discusses predictive equity and Chapter 5 describes DIF analyses for BMAT in detail, so they will not be revisited here.

Cambridge Assessment Admissions Testing recommends that BMAT results are used alongside other selection criteria in making admissions decisions. BMAT provides a measure of a student's ability to perform academically in pre-clinical course work; however other attributes, such as interpersonal skills or motivation to study medicine are also important, and frequently assessed in the admissions process. It is also acknowledged that medical schools need autonomy in determining the specific ways in which they use test scores in their admissions process. The way in which BMAT is used within the admissions process is largely determined by the policies and practices of the individual university departments that use it. Thus there are a number of ways in which BMAT is used. For example, some medical schools use 'cut-off' scores and will only consider applications above a minimum score. Others use BMAT scores to conduct an initial ranking to determine which application they will fully review first. Some medical schools use BMAT scores to determine who will be invited for interviews while others consider scores after interviews. Furthermore, some give equal consideration to all three sections, while others may give more importance to scores on a certain section (for example, BMAT Section 2).

The practices around test use are shaped by the particular needs and values of each university.

Due to the range of ways that universities use BMAT in their admissions process, Cambridge Assessment Admissions Testing supports a 'BMAT Liaison Group' in which universities share their admissions practices and discuss the issues they are facing. Cambridge Assessment hosts these twice-yearly meetings, to which representatives of all the faculties that use BMAT are invited. The meeting is an opportunity to update on recently completed research, to recap the key issues from the previous live session, and to explore questions around the nature of admissions test use. This forum also provides support for new institutions using BMAT for the first time. Admissions tutors from universities where BMAT has been used for a number of years are able to outline how the test fits into their own processes of selecting candidates for interview, or for an offer of a place, and can discuss the impact that the test has had on their own admissions rounds for the benefit of new users.

Understanding the ways in which test scores are used (and establishing that the use of scores from high-stakes tests in decision-making processes is justified) is an important aspect of consequential validity for stakeholder institutions. In Box 7.4, Professor David Vaux describes how the University of Oxford uses BMAT in conjunction with other indicators to shortlist candidates for interview, plus the way in which they assess the validity of the test for its intended purpose and its perceived value in their admissions process. This case study describes how Oxford's own monitoring of test use and candidate performance ensures that the actions arising from their use of BMAT scores are appropriate.

Box 7.4 Professor David Vaux[1] on the use of BMAT at the University of Oxford

BMAT was introduced for our undergraduate Medicine admissions in 2003. The primary use of the BMAT test when it was first introduced was as a shortlisting tool**.** There are far more applicants for Medicine than can be interviewed, so some method was needed for deciding whom to call for interview. In addition, very many Medicine applications are not from the UK, so comparisons have to be made across candidates in different school systems taking different school exams. For instance, in 2013 only 68% of candidates had GCSEs. BMAT is extremely useful in this context as a piece of data that is available for all candidates.

1 Professor David Vaux is Nuffield Research Fellow in Pathology and Tutor in Medicine at the University of Oxford.

How do we use BMAT?

BMAT is one component of the information used to generate a combined score for our initial algorithmic shortlisting process; the other component is a contextualised GCSE score where available. Approximately 90% of our shortlist is drawn from the top-ranked applicants using this combined score; all remaining applications are then inspected individually to ensure that all mitigating circumstances are appropriately taken into consideration, resulting in the addition of the final approximately 10% of our shortlist. The interviewers in each panel at both colleges do not know the BMAT score (or the college choice) during the interview process. College tutors receive BMAT scores and second college interview rankings only after they have submitted their own interview ranking. Tutors then make their final decisions based upon all of the separate items of information available to them.

Assessing the validity of BMAT as a selection tool

We carry out an annual analysis of the relationships between performance on indicators available during the selection process and performance during the course (separated into performance in the first and second year course, the Bachelor of Medicine (BM) examination, and overall performance in the third year Final Honour School (FHS) degree examinations.

Based upon recent comparisons across results for three years (2010, 2011, 2012), the statistically significant factors affecting BM examination performance are the total BMAT score, the mean interview score and gender. Only the total BMAT score and the mean interview score are effects that are stable over time.

There are two statistically significant factors that affect the FHS performance (the average score and the classified outcome) – the total BMAT score and the BM1 result. Both factors explain around 23% of the total variability of the FHS result suggesting they can be useful in predicting the FHS performance. Although the BM1 result is more useful and important in predicting FHS performance, this score is not available during the selection process. Of the information available during the admissions process, only the total BMAT score was shown to be a statistically significant predictor of academic performance.

An ongoing analysis of outcomes at the end of the 6-year standard medical course for the five cohorts for which data is now available suggests that there remain some statistically significant correlations between BMAT performance and performance in clinical finals (second BM examination) six years later, although this is a preliminary analysis and the effects are not seen for all sections of BMAT for all years. This is perhaps unsurprising, as BMAT is designed to assess academic skills that are, perhaps, more important in the pre-clinical years and less relevant in the clinical years, in which assessment of professional practice plays a greater role.

Stakeholder involvement
An important aspect of any test is the extent to which it retains the confidence of its users. Cambridge Assessment Admissions Testing has worked to ensure the engagement of stakeholders in a process of continuous scrutiny of the utility and performance of BMAT. In addition, Cambridge Assessment Admissions Testing has been pro-active in driving future development of BMAT, and ensuring that this evolution is directed by the needs of the stakeholders.

7.4 Is the washback effect positive?

BMAT is designed to both decrease negative washback (time spent on solely test-related knowledge or skills) and promote positive washback, by encouraging the development of academic skills relevant to success at medical school. BMAT is intended to require minimal preparation by students, and focuses on developing skills, such as problem solving, which will benefit them beyond the context of the test. The approaches taken to design positive washback and equitable access to test preparation materials are described in the following part of the chapter.

BMAT's explicit purpose is to help admissions tutors select the candidates with the potential to succeed in fast-paced, demanding, science-based courses (such as the non-clinical parts of medicine courses). This alignment of test content to medical course study is important for positive washback effects. BMAT does not aim to be a context-free measure of intelligence; rather, preparing for BMAT is directly related to school studies and future learning at university.

As discussed in Chapters 3 and 4, the content of BMAT Section 1 is not tied to any particular topics in school curricula and does not require specialised knowledge beyond basic computations for the problem solving items. BMAT Section 3 (Writing Task) also assumes no content knowledge. Nonetheless, the skills elicited by these sections are learnable skills, the practice of which is likely to have a positive effect on future learning.

BMAT Section 2 requires the application of scientific content knowledge. The content knowledge assumed by Section 2 is based on the National Curriculum for England and Wales and the GCSE science and maths specifications for the major UK examination boards. The syllabuses for international qualifications are also reviewed by Cambridge Assessment. As most candidates who are applying to study medicine or veterinary medicine will hold at least an A grade at GCSE (or equivalent) in two or three science subjects, and will be studying a combination of sciences for their A Levels, BMAT candidates should find that their preparation focuses on

revising and refreshing knowledge rather than learning large amounts of new material.

For mature students, preparing for BMAT Section 2 may require greater effort as applicants are likely to have lost familiarity with school-leaving science content at the point of applying to medical school. In this instance, preparing for Section 2 presents useful washback as applicants focus on learning, or relearning, science content that will be foundational to pre-clinical course work undertaken in medical education. Indeed, for some admissions tutors using BMAT in graduate-entry courses, the test's role in promoting positive washback is a key reason for using the test.

For high-stakes exams such as A Levels, which follow a specific curriculum with formal teaching input, support for teachers is central to fostering positive impact. For BMAT this is not the case: specific teaching and specialist preparation is not required. The concentration on core biology, chemistry, mathematics and physics in BMAT Section 2 (endorsed by admissions tutors as important to success as a medical student) means that any revision done will support and complement candidates' preparation for school-leaving exams, rather than divert their attention from their studies. Furthermore, the fact that BMAT questions do not rely on factual recall alone, but require knowledge to be applied and recombined in novel ways to reach solutions makes preparation for BMAT useful for encouraging thinking skills conducive to university-level study.

The timing of BMAT is further also intended to minimise negative washback. In the UK, BMAT usually takes place on the first Wednesday in November each year, and has traditionally been timetabled in order to fit with the universities' schedules for shortlisting and interviewing applicants to medicine. As the majority of BMAT candidates are in their final year of school study, it has been considered important that preparation for the test would not affect their usual school performance nor eat into valuable study time – a concern mitigated by a test date early in the academic year.

As described previously, it is unlikely that BMAT would influence the larger system of secondary education, and washback in this context primarily concerns self-directed test preparation activities that students undertake. Nonetheless, applicants to medical school spend a substantial amount of time in self-preparation for an admissions test, so the considerations for optimising test washback are important. The following part of the chapter outlines some of the key findings on the consequential validity of BMAT.

7.5 How are candidates preparing for the test?

Key research study – An investigation into candidates' preparation for the BioMedical Admissions Test (Gallacher, McElwee and Cheung 2017)

Main findings

- The majority of students feel well prepared for BMAT.
- Attempting practice tests under timed conditions is associated with achieving better test scores.
- There are some gender differences in feelings of preparedness, test preparation strategies and test outcomes.
- Commercial courses and extra help from schools are not associated with better test outcomes.

Introduction

As described earlier in the chapter, understanding the ways that students prepare for BMAT is important in gaining a picture of the wider test impact and washback effects of BMAT. The research summarised below investigated candidates' preparation for BMAT and how preparation strategies may influence test performance. These studies investigated the role of help received through commercial preparation courses, from schools and self-preparation activities, such as self-directed study with the free preparation materials provided by Cambridge Assessment Admissions Testing.

The research was designed to explore the strategies and materials used by students to prepare for BMAT, as these are the main washback effects of tests. Furthermore, it was hypothesised that preparation would be influenced by background variables, such as socio-economic status and school type, and that the amount of preparation help available to candidates may influence their test scores. Better-resourced schools, especially in the independent sector, are often better placed to devote time to helping candidates with special tuition and exam techniques. Students from independent schools are already over-represented in professions such as medicine (Milburn 2012). It was therefore deemed important to investigate whether (a) there was any evidence of systematic preparation in the independent sector, and (b) whether this translated into better test scores for this cohort, which would threaten the differential validity of the test and its equity for all candidates. However, estimating the impact of preparation from schools and from commercial coaching organisations is tricky, and some large-scale US studies using data from Scholastic Aptitude Tests (SATs) and the American College Test (ACT) have found that coaching gains have been largely over-estimated (Briggs 2001, 2004).

Claims made by commercial organisations offering test preparation that boosts scores need careful critical analysis. Firstly, estimates of score gains from commercial coaching must be made in relation to a control group of similar students who did not prepare for the test with a commercial programme – without this control group, any test preparation 'effect' is misleading. A second, perhaps more challenging, problem is that the groups of students who opt to pay for commercial coaching or not are not assigned at random. Those who choose to pay for coaching are actively self-selected and, as a group, may differ on other important variables also related to admissions test performance such as conscientiousness, motivation, or family encouragement characteristics.

In order to address issues on consequential validity related to test preparation, Cambridge Assessment produces BMAT preparation materials and makes these freely available. These materials include specimen and past papers, worked examples, answer keys and, recently, the *BMAT Section 2: Assumed Subject Knowledge guide*, which is a revision tool focused on the science knowledge needed for Section 2. Making these freely available is intended to provide all students with equal access.

Cambridge Assessment Admissions Testing maintains that its policy of making BMAT preparation materials available for free on its website means that commercial preparation courses are unnecessary. The website states that:

> Anyone offering a paid service to help you pass your admissions test(s) will have no more knowledge than someone who has read this website and studied past papers. So while a learner's performance at any test will improve with some familiarisation or practice, we would not advise candidates to pay for such help.[2]

Box 7.5 Preparation resources provided on the BMAT website[3]

BMAT Preparation Guide
Practice papers for Sections 1, 2 and 3
Worked examples for Sections 1, 2, and 3
BMAT Section 2: Assumed Subject Knowledge guide
Sample essays with examiner comments for Section 3
Test specification
Short videos introducing the test and on student experience with the test

The study presented here was conducted to inform Cambridge Assessment practices and policies, and to provide an evidence base for the preparation

2 www.admissionstestingservice.org/for-test-takers/preparation-materials
3 All resources on the BMAT website are openly accessible and free of charge.

guidance given to candidates. In addition, the research aimed to estimate the prevalence of preparation course use in the BMAT candidature. After presenting the research study, we describe how these research findings have informed the development of free resources that are made available to test candidates.

Study aims

The analysis presented here draws from two surveys that investigated candidates' preparation behaviours and their feelings of preparedness. Survey responses were also linked to BMAT scores where possible, to explore the relationships between self-reports of preparation and performance on the test. The research aims of this analysis were:

- to gain an understanding of the preparation behaviours of BMAT candidates, including use of help beyond the support freely available on the Cambridge Assessment website
- to test for relationships between preparation behaviours and BMAT performance, including the use of help beyond the support freely available on the Cambridge Assessment website
- to gain an understanding of the feelings of preparedness, and how useful each source of help is in preparing for BMAT.

Study methods

The main survey consisted of items about demographic background, feelings of preparedness, use of preparation materials and details of external help received from either schools or commercial organisations. Online delivery was used to administer the questions and the survey was made available on the BMAT website after candidates had already taken the test; approximately half of respondents responded before knowing their BMAT scores, while the rest responded after test results were released. Participation was voluntary and results presented here are anonymous, but candidate details were collected to enable matching to BMAT results data from the November 2015 session. In addition to data from the 2015 survey, responses from a similar survey administered in 2007 and 2008 are reported (Emery 2010b). Although the surveys included similar questions, the sampling procedures were different between the studies: Emery (2010b) only included candidates who had successfully gained entry into medical study whilst the 2015 survey sampled candidates soon after sitting the test, some of whom may not have gone on to study medicine. Therefore, the results of the two studies should not be directly compared. The discussion here focuses primarily on findings from the 2015 survey; however results from the earlier survey are reported to triangulate findings across the studies.

Missing data was excluded from analysis on an analysis-by-analysis basis, instead of including only that with full sets of responses. Therefore the

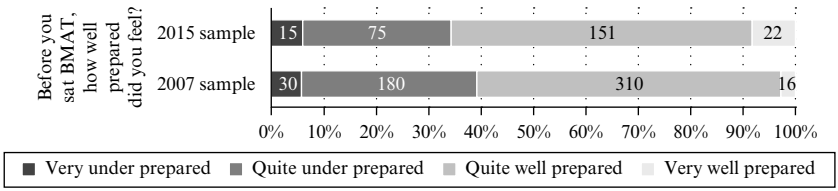
sample size 'n' varies considerably across analyses. This affects more complicated multivariate analyses more heavily than simple analyses, but allows maximal use of the data. The cohort is analysed as a whole, and also divided by gender and school type to investigate patterns of differential responses and scores.

Results

Feeling prepared for BMAT

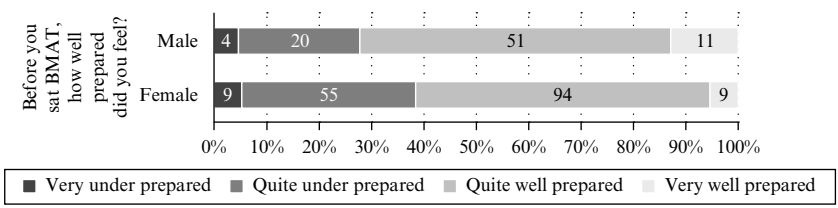
Candidates were asked to report how prepared they felt before taking the test. These perceptions of preparedness were investigated across groups by gender and school category. The majority of respondents felt 'very well prepared' or 'quite well prepared' (66%), and a low proportion of respondents felt 'very under prepared' (6%). Figure 7.3 shows the distribution of responses from the 2015 and 2007 samples side by side, showing that they are broadly similar across the two administrations, with a slight increase in the proportion of respondents who felt very or quite prepared.

Figure 7.3 Feelings of preparedness for the 2015 sample and 2007 sample



There were significant differences in feelings of preparedness across respondent backgrounds. Female respondents reported feeling less prepared than male respondents (Mann-Whitney U = 6165, $p = 0.038$), as can be seen in Figure 7.4: 72% of males felt very or quite well prepared compared to only 62% of females. Most of this difference is accounted for by the fact that males were more than twice as likely to report feeling very well prepared than females (12.8% versus 5.4%) Emery (2010b) found similar gender differences in feelings of preparedness in the 2007 sample.

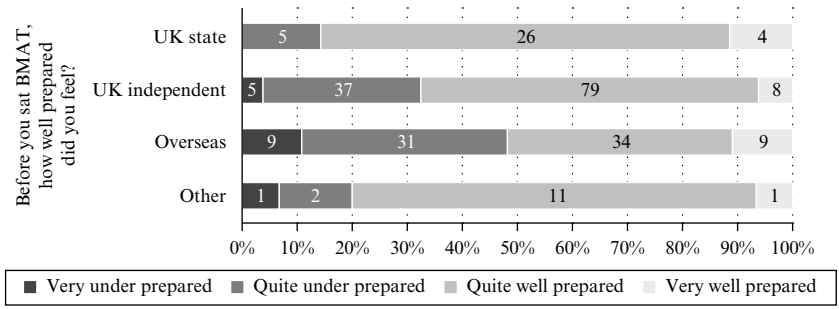
Figure 7.4 Feelings of preparedness for the 2015 sample, by gender



Whether this gender difference in self-reported preparedness is related to actual preparedness is unclear. Female respondents spent more time preparing for BMAT, on average, than males (see Figure 7.3), which might suggest that they were more prepared. The differences in reported preparedness might be related to gender differences in test anxiety or self-belief in their abilities. Female students score more highly on measures of test anxiety than males (Hembree 1988). Moreover, it is known that gender differences in self-belief persist, especially with respect to maths and science subjects at school. For instance, female students were found to be less confident of performing well on a maths test, despite negligible differences in actual test scores (Ross, Scott and Bruce 2012). This was found to be related to lower self-efficacy and a higher fear of failure in female students than in males.

It was found that there were significant differences in feelings of preparedness between respondents from different school types ($\chi^2 = 11.22$, $p = 0.011$). Despite the widely held assumption that students from independent schools receive more support in preparing for BMAT than state-schooled students, this was not borne out by the responses in the 2015 survey (Figure 7.5). UK state school respondents reported feeling better prepared than overseas students or respondents from independent schools. In contrast, the previous survey (Emery 2010b) found no association between school type and feelings of preparedness. In the intervening period, Cambridge Assessment Admissions Testing has improved the provision of preparation materials for test takers on their website, to include a *BMAT Section 2: Assumed Subject Knowledge guide*, explained answers for Sections 1 and 2 practice papers, and examples of Section 3 answers, with examiner comments. As these resources are freely available, state-schooled test takers may feel more reassured that they have been able to adequately prepare for BMAT, and this may be reflected in their responses in the 2015 survey.

Figure 7.5 Feelings of preparedness for the 2015 sample by school type



Across the entire 2015 survey sample, more hours of preparation was related to feeling more prepared generally. Moreover, those who felt more prepared were significantly more likely to perform better at BMAT Section 1 ($r_s = 0.164$, $p = 0.005$), Section 2 ($r_s = 0.164$, $p = 0.005$), and Section 3's quality of content ($r_s = 0.121$, $p = 0.031$) than respondents who felt under prepared. These effects of feelings of preparedness on test performance are small, but significant.

Participants were also likely to report feeling better prepared for BMAT if they had looked at the specimen tests on the BMAT website, had tried the specimen tests under timed conditions, or had used the BMAT preparation book. These ratings indicate that there are some very simple steps that candidates can take to feel better prepared for BMAT, which may positively impact their confidence on the test day.

Self-study for BMAT

Self-study was defined as using past papers, the *BMAT Section 2: Assumed Subject Knowledge guide*, and textbooks, but excluded time reported engaged in preparation support sessions delivered by schools or attending preparation courses. The median amount of self-study time reported for BMAT preparation was approximately 30 hours in the 2015 survey. This represents a considerable increase from the 2007 survey, when the median reported preparation time was eight hours (Emery 2010b). This could be a consequence of the additional preparation materials made available to test takers on the BMAT website. Out of 295 survey respondents, six respondents reported spending over 200 hours for BMAT overall (the maximum reported being 10,000 hours), so any value greater than 3 standard deviations from the median was treated as an outlier and excluded (greater than 216). Figure 7.6 displays the distribution of responses.

The time spent preparing for each BMAT section varied considerably (Table 7.2) with respondents spending much more time (on average) on Sections 1 and 2, than on Section 3. The difference between the time spent preparing for the different BMAT sections was more striking than in the previous survey (Emery 2010b): 2015 respondents spent (on average) 6-fold and 5-fold more time preparing for Sections 1 and 2, respectively, than the 2007 respondents and twice as long preparing for Section 3.

As mentioned above, female respondents reported spending more time preparing for BMAT (on average) than their male counterparts (Table 7.3). However, the gender differences in reported preparation time were not significant for Sections 1 and 2. This is in contrast to the previous study (Emery 2010b), which found that females reported a significantly greater number of hours preparing for BMAT Section 2 than males. From the 2015 sample, only the difference between genders in Section 3 preparation hours was significant (Mann-Whitney U = 5664.5, $p = 0.026$).

Figure 7.6 Hours spent preparing for the BMAT

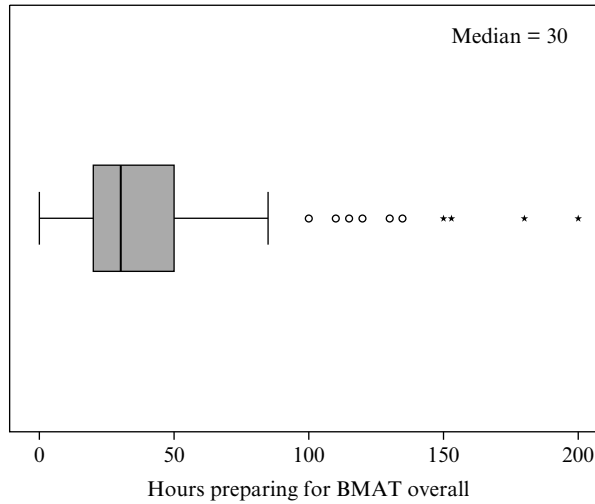


Table 7.2 Median numbers of hours' preparation for the 2015 sample and 2007 sample

Hours preparing for:	Median	
	2015 sample	2007 sample
BMAT overall	30	8
BMAT Section 1	12	2
BMAT Section 2	15	3
BMAT Section 3	4	2

Table 7.3 Median numbers of hours' preparation for the 2015 sample, by gender

Hours preparing for:	Median	
	Male	Female
BMAT overall	29	32
BMAT Section 1	10.5	11
BMAT Section 2	14	15
BMAT Section 3	3	5

Some interesting differences in the median number of hours spent preparing for BMAT were observed between different school types (Table 7.4). Respondents from the Other category (the majority of whom are 'mature' applicants over 21) reported spending the most time

preparing for BMAT, followed by respondents from UK independent schools. However, the difference was only significant with respect to the number of hours spent preparing for Section 3 (Kruskall-Wallis test $\chi^2 = 10.87$, $p = 0.012$).

Table 7.4 Median numbers of hours' preparation for the 2015 sample, by school category

Hours preparing for	Median			
	UK state	UK independent	Overseas	Other
BMAT overall	29	35	31	42
BMAT Section 1	11	15	10	15
BMAT Section 2	14	15	15	20
BMAT Section 3	4	5	2	10

Almost all respondents (95%) reported that they used the BMAT website for preparation (Figure 7.7) and about 90% looked at the full specimen tests for Sections 1 and 2; whereas only 80% of respondents had used the Section 3 specimen paper (Figure 7.8).

Figure 7.7 Sources of help while preparing for BMAT 2015

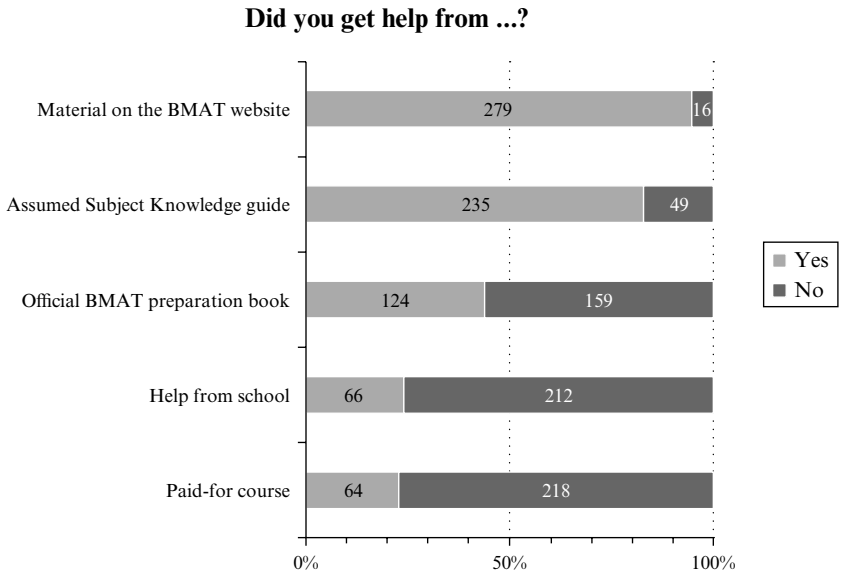
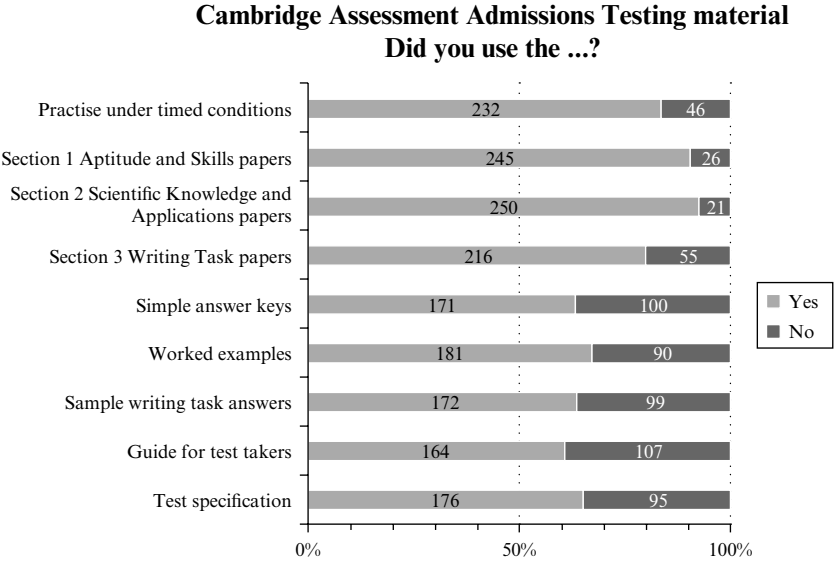


Figure 7.8 Use of Cambridge Assessment Admissions Testing material while preparing for BMAT by the 2015 sample

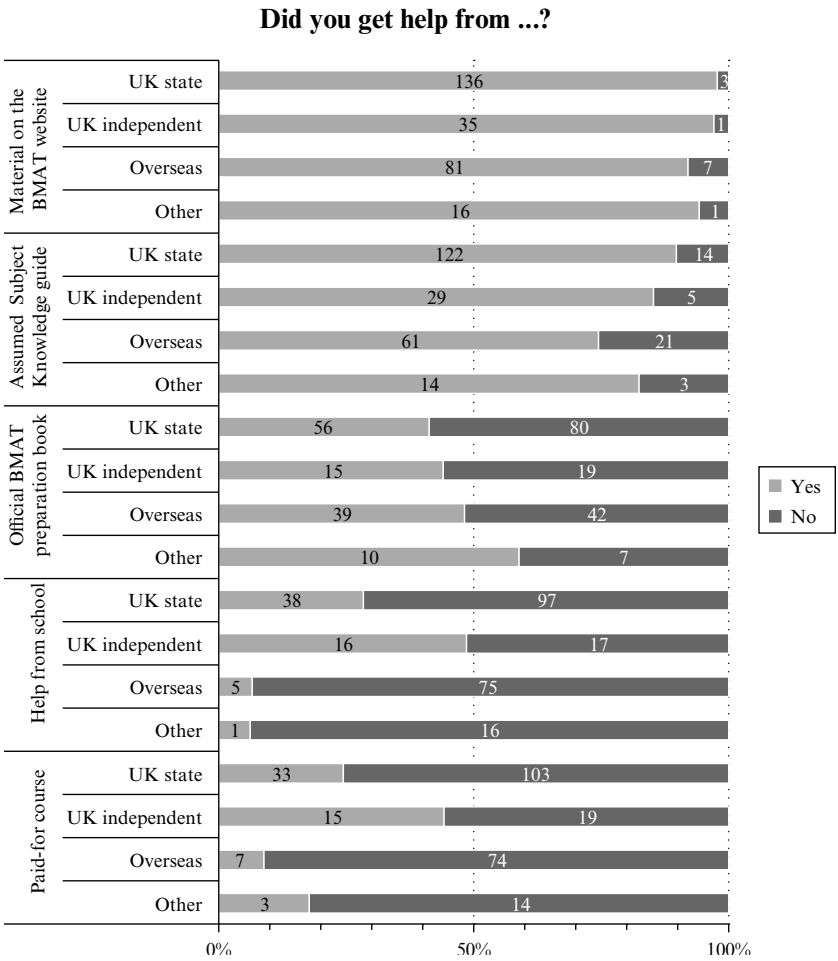


In the 2015 sample, the majority of candidates (83%) who looked at practice test papers reported practising under timed conditions (Figure 7.8). This suggests a change in test preparation behaviours, as Emery (2010b) found that only one third of students who reported using the specimen papers had practised them under timed conditions. Females reported lower rates of practising under timed conditions than did males. The reasons for this are unclear but could contribute to female respondents feeling less prepared for the test than males (see Figure 7.4).

External help preparing for BMAT

While the free materials provided by Cambridge Assessment Admissions Testing were widely used, only a minority of students in 2015 reported receiving help from their school or from commercial courses: overall, 24% of respondents reported getting help from their school and 23% reported attending a course (Figure 7.7). However, there was considerable difference between candidates from different school backgrounds in the likelihood of receiving external help. Overseas respondents were the least likely to access external help in preparing for BMAT, followed by those from the 'other' category. In common with the findings from the previous survey (Emery 2010b), candidates from UK independent schools were much more likely to receive help from their school, or attend a course, than respondents from other

Figure 7.9 Sources of help while preparing for BMAT, by school type



backgrounds. Almost half (48%) of UK independent school respondents had received help from their school, compared to 28% for UK state schools; and 6% each for overseas and 'other' (Figure 7.9). This may not be surprising, in light of other evidence that independent schools invest more time and effort in preparing their students for applying to medical school (Wright 2015). Similarly 44% of UK independent school respondents had attended a commercial course, compared to 24%, 9% and 18% for UK state schools, overseas and 'other' respectively (Figure 7.9). Again, this is perhaps unsurprising: parents who pay for their children's education are more likely to be able to afford to pay for a commercial course. The fact that independently schooled

respondents were much more likely to access external help in preparing for BMAT could be perceived as affording them an additional advantage in applying to medical school. This could be a cause for concern for medical schools and candidates from outside of the UK independent school sector, but it should be noted that the results from the 2015 survey do not demonstrate any association between accessing these forms of external help and improved BMAT scores (see the next section).

The 2007 survey also found that there was considerable variation between different school backgrounds in the amount of school help accessed: for those students who *did* report receiving school help, the modal (most frequently reported) number of hours of help accessed was five to nine hours for independent school students, three to four hours for grammar/selective school students, and one hour for comprehensive school students (Emery 2010b). In this survey, the majority of respondents reported receiving no preparation help from their school. For those that did report receiving school help, this amounted to an average of three hours (which was the same for state and independent school students) and was most frequently in the form of advice on BMAT test contents rather than organised tuition/revision classes.

Relationships between types of preparation and BMAT scores

Relationships between preparation behaviours and BMAT performance were tested using correlations and hierarchical regression analysis. In each analysis, the impact of gender and school type was investigated, as well as the self-reports of test preparation behaviours from the survey.

When interpreting these findings it must be considered that there is no baseline measure of candidates' ability, and so it is unknown whether candidates choosing different methods of BMAT preparation were of equal ability at the outset. Choice of preparation method may be related to other characteristics that determine test performance that cannot be controlled for without additional data. *Causal* relationships between preparation methods and test performance should therefore not be inferred.

Bi-variate analysis shows that attempting practice tests under timed conditions was associated with higher tests scores on all sections of BMAT. In contrast, attempting practice tests without time constraints was associated with poorer test performance. Similarly, looking at practice papers without attempting to answer them was also associated with poorer performance on Sections 1 and 2. This finding was confirmed through multi-variate analysis; practising under timed conditions remained a significant predictor of test performance on all three sections of BMAT when controlling for other factors.

None of the following preparation behaviours was found to predict test performance in multi-variate models controlling for other factors: using materials from the BMAT website, accessing help from school or attending a

commercial course. Moreover, no association was found between the hours spent preparing for BMAT and test scores.

Differential validity was also investigated. After controlling for test preparation behaviours, gender (male) was significantly associated with better performance on Section 2, though not on Sections 1 or 3. Issues relating to gender differences in test performance have been described in Chapter 2. In contrast, no association was observed between the following factors and better test performance: school type, help received at school or attending commercial preparation courses. This is encouraging in terms of social justice and access to the medical profession for state-schooled students from lower income backgrounds, who may attend schools that are less experienced in preparing their students to apply to medical school (Wright 2015) and who cannot afford to pay for a preparatory course.

Discussion

This study into BMAT preparation investigated consequential validity questions about how candidates are preparing for BMAT, feeling of preparedness and associations between preparation strategies and test performance. Throughout the analysis we investigated whether there were differential effects by gender or school type.

Candidates now spend 30 hours (on average) preparing for BMAT, a substantial increase from the reported average of eight hours in the 2007 survey (Emery 2010b). The reasons for this increase are not known but one possible explanation may be that there are now considerably more resources freely available on the BMAT website than there were at the time of the previous survey. As candidates are investing a substantial amount of time on test preparation, the issue of washback is particularly important to ensure that time spent on test preparation has educational value beyond performance on the test.

Attempting tests under timed conditions is associated with better test performance, and based on this finding the test preparation guidelines on the BMAT website encourage students to use this technique. Simply looking at papers without attempting them, or attempting papers without time constraints were both associated with poorer performance, suggesting that these are ineffective or even counterproductive preparation behaviours. If candidates become used to spending more time than realistic per question, they will be less able to answer the questions in the time available during the live administration, causing undue stress and underperformance compared to practice papers. We found that females are less likely to practise under timed conditions, and tend to have slightly lower scores on Section 2; however, we cannot infer causation between these findings.

While a quarter of students receive external help from either schools or commercial courses in test preparation, there was no evidence that this resulted in higher test scores when controlling for other variables. From this

we conclude that while students from different socio-economic backgrounds engage in different test preparation strategies, there is not evidence that this systematically gives any group an advantage on test performance.

This study aimed to provide a picture of how candidates prepare for BMAT and demonstrated that there are some simple, and low-cost, ways to prepare for the test that impact upon candidates' sense of test-readiness. Any gains from commercial coaching, while difficult to estimate in a correlational design, were not apparent from this data.

7.6 How is the test perceived by stakeholders?

Key research study – Student perceptions of the medical admissions process (Emery and McElwee 2014)

Main findings

- Perceptions of admissions tests are not a deterrent to applying to medical study.
- Admissions tests are seen as 'daunting' for similar reasons as interviews.
- There are gender differences in how admissions test are perceived.

Introduction and context

A key group of BMAT stakeholders are the test candidates themselves – it is important to investigate whether candidates view BMAT as fair and whether their perceptions of the test pose a barrier to applying to university. A piece of research was carried out (Emery and McElwee 2014) to investigate candidates' perceptions of admissions tests within the wider context of the medical applications process.

Selection to study medicine in the UK is an area of particular challenge with respect to widening participation (that is, the desire to increase the proportion of students in higher education who come from traditionally under-represented – i.e. more socially disadvantaged – groups). A much higher proportion of students of medicine and dentistry in the UK come from the higher socio-economic groups (Steven, Dowell, Jackson and Guthrie 2016). In the late 1990s, students from social class I (whose parents have professional occupations) were 30 times more likely to gain a place at medical school than those from class V (whose parents have partly skilled or unskilled occupations) (Seyan, Greenhalgh and Dorling 2004). More recently, it was found that applicants from social class I still predominate and those from lower social classes are significantly under-represented in the applicant pool. Moreover, those from National Statistics Socio-economic Classification

(NS-SEC) class 1 are more likely to be successful in their application than those from the lowest social class (NS-SEC class V) (Steven et al 2016).

Widening participation (WP) in medicine has particularly difficult challenges to overcome. Entry requirements are competitive, with admissions tutors generally looking for three A grades at A Level as a minimum. This is an issue as A Level attainment is closely related to social class and it is acknowledged that candidates from WP backgrounds may have school grades that underestimate their potential for higher education study (Hoare and Johnston 2011). Applicants to medicine are also generally required to demonstrate that they have acquired work experience (usually unpaid) in a medical or community setting. This may also disadvantage students from WP backgrounds for various reasons: because they cannot afford to do voluntary work; public transport links are not suitable if they are from rural areas; or they have lesser access to personal or family connections through which they can organise suitable placements. Finally, the length of the degree course in medicine may be a deterrent in respect of the higher tuition fees and associated costs of accommodation.

With all of these competing issues, it is important that an admissions test such as BMAT does not pose an additional barrier to entry for students from WP backgrounds (or from state schools, who are also under-represented in medical study). Emery and McElwee's (2014) study aimed to take a 'student voice' perspective and explore how potential medicine applicants perceive admissions tests such as BMAT, as part of the applications process for medical study in the UK.

Research questions

1. Do students' views on admissions tests and other selection criteria for medical study differ according to their social and educational backgrounds?
2. Are students from WP backgrounds (or from state schools) more likely to view admissions tests as a deterrent to applying to study medicine?

Data collection and analyses

This study used a convergent mixed methods survey design to investigate students' perceptions of selection methods (academic achievement, admissions tests, traditional interviews and MMIs) and potential differences according to gender and socio-economic status. A survey including demographic items, Likert-rating scales and open-ended questions was distributed to students interested in applying to medical school. The survey was distributed at medical school 'open days' for perspective applicants, and on the BMAT website. Participation was voluntary.

WP indicators were collected; including whether the respondent was eligible for free school meals or education maintenance bursaries, or whether

they were the first in their family to attend university. Students from state schools who also had met one of the WP indicators were considered as potentially having WP status for medical school admissions and were classified as WP in our analysis.

Questions on attitudes towards the selection criteria were mostly in Likert-scale format (ratings on a 5-point scale from 1 'not at all' to 5 'very'). These questions asked students about their perceptions of the fairness, usefulness and relevance of each of the criteria, how daunting they considered each to be, and their level of confidence that they could perform well on these. Open-ended questions were included for students to expand on their responses to Likert-scale questions.

Likert responses were analysed to calculate mean ratings for each selection method and these were used to investigate differences between groups. Responses to open-ended questions were analysed through a general inductive approach. In a second stage, quantitative and qualitative results were integrated and interpreted leading to a second round of qualitative analysis.

Throughout the survey, the questions referred to 'admissions tests' (or specific admissions test skills) rather than specifically asking about BMAT. A number of skills tested by the United Kingdom Clinical Aptitude Test (UKCAT, www.ukcat.ac.uk), another widely used admissions test for medicine in the UK, were also included to gain a more rounded picture of students' views overall. The selection criteria investigated were:

- GCSE grades
- A Level predicted grades
- personal statements
- teacher references
- relevant work experience
- admissions tests in general†
- admissions tests – verbal and numeric reasoning skills†
- admissions tests – abstract reasoning skills
- admissions tests – subject-specific reasoning skills†
- admissions tests – writing skills†
- admissions tests – behavioural skills
- traditional interviews
- MMIs.

A brief definition of each selection criterion was provided to ensure that all respondents understood what each element involved. The questionnaire was piloted and then distributed at open days at BMAT institutions, at an outreach summer school, and electronically via the BMAT website. In total,

† These are sections included in BMAT.

data from 749 respondents (63% female, 37% male) who indicated that they were considering applying to study medicine in the UK was included for analysis. Almost 55% indicated that they were in the second year of their A Levels (or equivalent), with the rest of the responses from candidates in the first year of A Levels, or with their secondary school education completed. Approximately 80% of respondents were based in the UK.

Results

Of the UK respondents, 79% were categorised as attending a state school and 21% as attending an independent (fee-paying) school. Of the two thirds of the sample who provided sufficient background data to enable classification, 39% were identified as potentially meeting WP criteria. The results below are organised according to specific questions from the questionnaire.

How daunting/scary do you think you would find the following admissions criteria?

Mean participant ratings for this question ranged from 2.07 (for relevant work experience) up to 3.89 (for admissions tests in general). All criteria, apart from relevant work experience, received median ratings of 3 or 4, suggesting that students find the applications process overall quite daunting. It is interesting to note that, although admissions tests in general received quite a high rating, the ratings for the individual admissions test skills were lower. Of the specific admissions test skills, abstract reasoning skills received the highest mean 'dauntingness' rating (3.51), with verbal and numeric reasoning skills and subject-specific reasoning skills, as tested in BMAT, receiving lower ratings (3.37 and 3.06 respectively). Students did report finding writing skills, also in BMAT, somewhat more daunting (mean rating of 3.40). Males reported finding every aspect of the selection process less daunting than did females. By contrast, there was no difference in ratings by students from state and independent schools, nor according to whether they were classified as from a WP background or not.

Qualitative analysis revealed that students find admissions tests, traditional interviews and MMIs daunting for similar reasons. These selection criteria are 'one-off' chances and are seen as challenging. Students worry about not performing to their 'true ability' due to test anxiety, illness or an unusually poor performance. Furthermore, interviews and admissions tests are seen as final 'hurdles' that must be cleared to gain entry to medical study. These issues are illustrated by the following quotes from applicants:

> Traditional interviews are very daunting and the fear of not portraying yourself well is always on your mind, and that your life could be changed by those 10 minutes.
> (Male student from a state school, first in family to attend university and eligible for free school meals)

> I feel that it's not fair, as you have achieved the grades, sorted your personal statement and work experience but fall at the last hurdle.
> (Male student from a sixth form college)

> These admissions tests are based on a one-chance day, and underperformance on one day could leave someone unable to get into their desired university, despite them having an excellent academic background and all the relevant work experience.
> (Female student from a comprehensive sixth form college)

How likely is it that the following admissions criteria would deter you from applying to study medicine at a particular institution?

Participants' ratings of the extent to which the various admissions criteria were a deterrent to application formed an interesting counterpoint to their ratings of how daunting they felt these to be. The modal rating for all criteria in this question was '1' (not at all). The picture that emerged from the data is that students may find the admissions process daunting but they are committed to the idea of applying to study medicine regardless. The qualitative comments supported this idea – when asked to comment on why certain aspects of the process might deter them, a significant proportion of those who commented remarked that nothing would deter them from applying.

Females rated the verbal and numeric reasoning, and the subject-specific aspects of admissions tests as slightly more of a deterrent than did males. However, the mean ratings were low for both genders. According to their ratings, students from state schools or from WP backgrounds are no more likely to be deterred by admissions tests than those from independent schools or non-WP backgrounds.

How fair is it to compare students from different educational and social backgrounds on each of the following admissions criteria?

Overall, respondents perceived all admissions criteria to be somewhere between 'somewhat fair' (score of 3) and 'fair' (score of 4). On average, none of the criteria was perceived to be 'unfair' (2) or 'not at all fair' (1). The lowest mean fairness ratings were given to abstract reasoning skills (3.22), relevant work experience (3.24), and writing skills (3.25). Verbal and numerical reasoning skills, and subject-specific reasoning skills (as tested in BMAT Sections 1 and 2), received slightly higher fairness ratings of 3.35 and 3.51 respectively. Perhaps somewhat surprisingly, participants gave the highest mean fairness ratings to traditional interviews (3.92). This is in contrast to the published evidence that suggests that traditional interviews are not as effective or as fair as some other selection methods (Patterson et al 2016). However, participants who were classified as WP rated traditional interviews as slightly less fair than did candidates who were classified as non-WP.

Qualitative analysis revealed that students acutely perceive the social and educational inequities at play in the admissions process and that these produce perceptions of unfairness. This was often described as unequal access to resources, including financial resources (e.g. to be able to pay for a commercial course), educational resources (what type of school they attend) and cultural capital (the social networks and knowledge that students access from their friends and families). The following quote illustrates these issues:

> Whilst this is virtually impossible to resolve, there is a massive social bias towards wealthier, better educated candidates. This is particularly significant in assessing candidates on their relevant work experience, which is easiest to acquire if an applicant has contacts in the medical profession. I also believe that . . . individualism is quelled by advice given to candidates over personal statements and interview responses.
> (Male student from a state sixth form school)

How confident are you that you could perform well on the admissions criteria?

Overall, respondents expressed a relatively high level of confidence that they could perform well on every criterion. Mean participant ratings for the 'perform well' question ranged from 3.34 (for abstract reasoning skills) to 3.96 (for personal statements). The mean rating for traditional interviews was slightly higher than those for MMIs and admissions tests in general. Of the specific admissions test sections, behavioural skills and subject-specific skills received the highest mean ratings, followed by verbal and numeric reasoning skills and writing skills and, lastly, abstract reasoning skills. Females were less confident in their capacity to perform well than males – although both genders still gave relatively high confidence ratings. State school and WP respondents were slightly (but statistically significantly) less confident that they could perform well on admissions tests overall than were independent school and non-WP respondents; however, when their ratings for individual admissions test sections/skills were examined there was no significant difference between the groups.

In your opinion, to what extent is help and preparation from other sources (e.g. school, tutors, parents, preparation courses) likely to have a large impact on performance on the following admissions criteria?

Personal statements were rated as the aspect of the admissions process most likely to be influenced by external help (mean rating of 4.12), followed by interviews (3.80), MMIs (3.59) and securing work experience (3.58). All these criteria received higher ratings than admissions tests. Of the specific admissions test sections, subject-specific skills received the highest mean rating (3.45), followed by writing skills (3.37), verbal and numeric reasoning skills (3.23), abstract reasoning skills (3.13) and, lastly, behavioural skills (3.04).

A forced-choice question was also posed to respondents: 'If you had a choice, in addition to academic performance (e.g. GCSE and A Level grades), on which criteria would you prefer your medicine application to be considered?' An interesting school-type trend emerged in the responses, although it did not prove to be statistically significant. A greater proportion of state school students stated a preference to have their application considered on the basis of both interview performance and admissions test performance (53% of state school students, compared with 43% of independent school students), while students from private schools were more likely to choose interview performance only (43%).

Discussion

This study examined the views of potential medicine applicants towards the various admissions criteria that might be considered as part of their application process. Questions did not canvass views on specific admissions tests but rather focused on the skills assessed within these tests.

Views towards admissions tests, traditional interviews and MMIs were mixed, with students generally finding these selection criteria 'daunting' but fair. In assessing the impact of admissions tests on the process of selection for medicine it is particularly important to contrast responses to the question of *how daunting* certain aspects of the selection process seem to the question of *how much of a deterrent* to application those same aspects present. While admissions tests in general were rated as the most daunting of the criteria listed, their rating as a deterrent to application was low. In fact, the modal rating for this question was a '1' (i.e. 'not at all' a deterrent) for all criteria. The majority of respondents stated that they were 'quite sure' or 'very sure' that they wanted to study medicine. Thus a picture emerges of candidates who are committed to the idea of studying medicine and, while they may find aspects of the selection process daunting, they are not deterred from applying in pursuit of their ambition. Students from WP backgrounds and from state schools did not report finding the prospect of an admissions test to be more of a deterrent than did the non-WP and independent school respondents.

Overall, the results of this study suggest that there are very few differences in how students from different social and educational backgrounds view the admissions process to medicine and admissions tests, in particular. More pronounced were some of the gender differences: females rated most criteria as more daunting than did males. Females also made a greater number of qualitative comments about the competitiveness of the application process. For all the criteria listed, males rated higher confidence than females that they could perform well. This is of particular interest given the small but stable gender differences in BMAT performance, discussed in Chapter 2, that do not appear to stem from any discernible test bias.

What do these results contribute to our understanding of the consequential validity of admissions tests? For medical school applicants, who are determined to get into medical school regardless of the obstacles placed in their path, the importance of an admissions test that impacts on the success or failure of their application is self-evident. Candidates reported finding admissions tests relatively daunting, which is likely to be related to how difficult they perceive the test to be, but also reported feeling relatively confident of performing well. All of these findings are consistent with elements of positive washback, and therefore consequential validity, as described by Green (2013): that the test should be perceived to be both important and difficult (but attainable). It is important to acknowledge the limitations of this research: the timing of the questionnaire in the admissions cycle meant that a number of respondents had already taken an admissions test other than BMAT, which may have impacted their views of BMAT. Further, the fact that the questionnaire was distributed at university open days and on the BMAT website means that early prospective applicants who were truly deterred from applying to medicine would not have been included; reaching this particular group is difficult. Nonetheless, this study provides important insights into students' views of applying to medicine and the place of admissions tests within that process.

7.7 Chapter summary

Consequential validity is an element of test validation that is critical to the fitness for purpose of any assessment. To explore the consequential validity of BMAT we have reviewed the practices of test design, stakeholder engagement and empirical research using a socio-cognitive validity framework (O'Sullivan and Weir 2011) and addressed issues of washback, test score use, test preparation practices, perceptions of the test and differential validity. As consequential validity can only be established after a test has been developed and used, the principles of *impact by design* aim to anticipate outcomes and mitigate possible negative effects at the test development phase.

By conducting research on the consequences (both perceived and real) of test use, test developers can seek to understand the impact of decisions that result from a testing policy. In medical education research, consequential validity focuses on the decisions made by tutors about how to interpret test scores (Downing 2003). However, we have shown that the decisions made by test takers, medical students and prospective test takers can also be considered as part of consequential validity. In particular, it was noted that consequential validity could impact future admissions cycles and decisions of potential applicants about whether to apply for biomedical study. Our findings indicate that applicants to medical school view forms of selection as relatively daunting, but these methods do not pose a barrier to applying (Emery and McElwee 2014).

We have presented research into the preparation behaviours and test perceptions of different groups of applicants by gender and school status, and this is an important step in fully investigating the differential validity of BMAT. However, there is another group of applicants – mature students – that has not been as rigorously investigated. Due to the different backgrounds of this group of students, and efforts from medical schools to enrol mature students, further research is needed in this area.

Through monitoring test use and maintaining effective collaboration with stakeholders, consequential validity can feed into test or curriculum revisions. If test providers and stakeholders understand how students prepare for a test, they can also adopt measures designed to support positive test impact, such as those that encourage test preparation behaviours designed to have a beneficial educational impact. This chapter has described how Cambridge Assessment Admissions Testing seeks to ensure the positive impact of BMAT in these ways, by providing enhanced support materials (such as the Section 2 revision guide), through focused research with past and potential candidates, and by encouraging dialogue with (and between) stakeholder institutions.

Chapter 7 main points

- Washback impacts on education systems differently for admissions tests when compared to language tests.
- Practising tests under timed conditions is associated with higher test scores on all sections of BMAT, whereas school type and attending a course are not.
- Perceptions of admissions tests are not a deterrent to applying to medical study. Students are committed to the idea of applying regardless.
- Further work on consequential validity should investigate how the social consequences of test use interact with other aspects of validity.
- The social consequences of assessment should be considered as part of test design.

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.