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

8 Conclusions and recommendations

Kevin Y F Cheung

Research and Thought Leadership Group, Cambridge Assessment Admissions Testing

Introduction

The socio-cognitive framework of test validity originally outlined by Weir (2005) has served as a springboard to investigate the examination of writing (Shaw and Weir 2007), reading (Khalifa and Weir 2009), speaking (Taylor 2011) and listening (Geranpayeh and Taylor (Eds) 2013). These previous volumes have contributed to Cambridge English Language Assessment's approaches to test development and revision by comprehensively evaluating Cambridge English examinations. Similarly, the present volume represents a stock-taking of Cambridge Assessment Admissions Testing's approach to assessment, focusing on the potential for biomedical study as conceptualised in the BioMedical Admissions Test (BMAT). The issues identified here serve as a focal point for revising admissions tests in the future, and development of tests for other contexts.

Although the socio-cognitive framework was originally developed to evaluate language tests, Weir pointed out that the model would be useful in other fields of educational assessment: 'Though specifically framed with English for Speakers of Other Languages (ESOL) in mind, the blueprint has implications for all forms of educational assessment' (2005:2). In the present volume, we apply the framework to the admissions testing context. This extends use of the socio-cognitive framework outside of the language testing domain, but we are admittedly not the first to do so. According to O'Sullivan and Weir (2011), the socio-cognitive framework has been applied to examinations assessing art, physics and ophthalmology, due to its usefulness for guiding discussions of validity. However, to my knowledge, this volume represents the most comprehensive application of Weir's socio-cognitive framework to an assessment setting outside of the language testing domain.

As one might expect, some issues relating to the validity of tests are different in the admissions testing context when compared with language testing. Notably, cognitive validity is particularly complex in admissions testing, due to the range of constructs that are plausible to assess in this context. Another

way that the field of admissions testing diverges from language testing is the focus on prediction when investigating criterion-related validity. Also, the concepts of consequential validity and washback for admissions testing stakeholders are different from the topics commonly explored in language learning. These differences have not been barriers to applying the socio-cognitive framework; instead, they have highlighted that the aspects of validity identified by Weir (2005) can manifest in various ways, and that these aspects are important to consider across all assessments. Perhaps more surprisingly for readers, there are numerous areas where the issues in language testing and admissions testing are similar.

The chapters of this volume have highlighted one important parallel between language testing and admissions testing – the advantage of adopting holistic perspectives when evaluating tests and their use. The range of topics covered by authors for this volume has vanquished the myth of the validated admissions test, by showing that the test itself is one part of a much larger context that responsible test providers must consider. Only focusing on the test in isolation could result in claims about the assessment that are not defensible once the situation surrounding the test administration is taken into account. By assuming that tasks assess relevant cognitive processes or ignoring the testing context, test developers can risk unintended consequences arising from introduction of an assessment, particularly one that is used for high-stakes purposes. In this regard, an approach to evaluating the entire testing policy has been adopted throughout this volume, as advocated by Newton and Shaw (2014). Unlike Newton and Shaw, however, we propose that an existing framework of validity, Weir's (2005) socio-cognitive model, is a sufficient starting point for this approach, as it already extends evaluation of validity beyond the technical aspects of a test.

Many of Weir's (2005) ideas regarding validity and language testing can be applied appropriately to admissions tests; however, there is one place where the Cambridge Assessment Admissions Testing position diverges from the perspective adopted by Weir. He argues that 'practicality is simply not a necessary condition for validity' (2005:49). Whilst I agree that the test provider must focus on the construct to ensure that practicality does not intrude and distort what we are aiming to assess, validity lies in the appropriateness of inferences made using the assessment, and practical issues can impact on these. For example, if universities do not receive results of an admissions test within a timeframe that supports their shortlisting decisions, the validity of the test is compromised. Therefore, the practical aspects of marking and returning results must be considered as part of validity, and we have included them in Chapters 4 and 5 on context validity and scoring validity. Similarly, the cost of producing and marking an admissions test must not make registration prohibitively expensive for candidates, as this would impact on interpreting results where the self-selected candidate pool has been unduly shaped

by factors not relevant to the test constructs, such as socio-economic status. These issues are apparent in the context of selection to study medicine, where widening access to higher education is emphasised. Many of these topics are touched upon in Chapter 7's exploration of consequential validity.

Notwithstanding the divergent views on practicality, the arguments made by Weir (2005) about language testing are remarkably similar to those presented throughout this volume on an admissions testing context. This applies to the current approach adopted when developing admissions tests and also to earlier work, particularly in relation to BMAT. The history behind various aspects of BMAT's validity has been presented in this volume and this represents a snapshot of a moment in the lifetime of the test. There are currently 17 universities in the UK and internationally who use BMAT for admission to more than 25 courses of medical, biomedical or dentistry study. This number is steadily growing and it is likely that new developments in the administration, delivery and scoring of BMAT will emerge in the coming years, as it serves an increasingly global higher education arena. Further challenges are potentially on the horizon that will need to be addressed with an evidence-based approach that considers all the aspects of validity identified in the socio-cognitive framework.

The rest of this final chapter turns to each aspect of the socio-cognitive framework to summarise the validity of BMAT viewed through the lens of Weir's (2005) model. Importantly, these summaries also identify areas for future research that can support investigation of validity going forward. Validity exists on a continuum and should not be regarded as a binary concept (Messick 1989); therefore, it is important to acknowledge that continuing efforts are needed to ensure BMAT's fitness for purpose.

Test taker characteristics

Cambridge Assessment Admissions Testing routinely monitors the test taker population and their performance on the three sections of BMAT. This approach acknowledges that the test taker is at the heart of the assessment and that test development should recognise the physiological, psychological and experiential issues that can impact performance. An understanding of the test taker population is important for considering all aspects of validity identified in the socio-cognitive framework. From a quality assurance perspective, information about test takers' gender and school background is used to check for bias in test items. In the context of BMAT, the predictive equity of test scores for different groups is an issue to consider and continue investigating, particularly as the population taking BMAT changes. The authors of Chapter 2 highlight the need to understand shifting educational contexts to guide this work going forward. As changes to education policy can influence the ways that certain groups are categorised or focused upon,

the practical issues involved in classifying students should be carefully considered for future work. To support this, engagement with the users of an admissions test is crucial. Many of Chapter 2's more nuanced observations about proxies for socio-economic status came from Brenda Cross, whose experiences as a seasoned medical school admissions tutor revealed the care that medical schools take when selecting applicants, and the complex array of considerations that they face. Admissions tests are always situated in a wider selection process that can include access arrangements and influence from government education policies. These issues are easily missed by a test developer without the input of users who are actually making selection decisions, and Cambridge Assessment's approach recognises the need to engage with admissions tutors as part of understanding the test taker.

There are avenues for research on BMAT's test taker characteristics that would contribute to literature outside the admissions testing domain. Linking Cambridge Assessment data to other sources, such as UCAS data, could be useful for understanding group differences, not just in BMAT performance, but also in the choices made by school leavers and applicants to medical school. Some research on university choice indicates that there are complex relationships between gender, distance of the institution from home and A Level choice (Gill, Vidal Rodeiro and Zanini 2015). Monitoring of the test taker population is also crucial for evaluating the performances of an international candidature with a diverse educational background. Over the last five years, an increasing number of medical schools have recognised the attributes assessed by BMAT as important and decided to include the test as part of their selection procedures. Departments in the Netherlands, Spain, Malaysia, Thailand and Singapore now require BMAT to be taken as part of the selection process. Universities in other countries are also at various stages of trialling and evaluating how BMAT fits into their procedures and policies. Monitoring the test taker characteristics of future sessions will contribute to understanding the specific challenges of assessing candidates from different education systems. Furthermore, an understanding of how international group performance interacts with more traditional group differences, such as gender, will be crucial to ensuring that BMAT remains fit for purpose.

Cognitive validity

Understanding the cognitive processes elicited by BMAT tasks is crucial to investigating the validity of the test. As Weir and Taylor (2011:299) point out: 'It is hard to see how one can build a convincing validity argument for any assessment practice without assigning cognitive processing a central place within that argument.' Suitable interpretation of test scores relies on extrapolating from performance on test tasks to real-world behaviours. Therefore, BMAT should elicit the kinds of mental operations that are relevant for

biomedical study. Chapter 3 uses cognitive validity as conceptualised in the socio-cognitive framework to present a key question for developers of admissions tests: what are the skills and cognitive processes that a test should aim to elicit and assess?

In the case of BMAT, this question is answered by presenting the rationale for assessing the skills targeted by the test, and the theoretical basis for conceptualising each skill as potential for biomedical study. The rationales for assessing generic thinking skills, scientific knowledge and application, and written argument were considered when designing BMAT as a successor to two earlier tests used for selection to medical study. The Oxford Medical Admissions Test (OMAT) and the Medical and Veterinary Admissions Test (MVAT) were used to select undergraduate students and deal with increasingly large pools of applicants. Both of these tests were designed to assess specific abilities theorised as important in biomedical study (James and Hawkins 2004, Massey 2004), which was identified as the real-world situation relevant for evaluating a biomedical admissions test. In Chapter 3, the original rationales were re-examined in the context of contemporary understandings of biomedical study. Although a wide range of topics are included in biomedical study, various sources agree that core skills are relevant for biomedical study. Trainee clinicians are engaged in rigorous learning and need to develop problem solving skills (Quality Assurance Agency 2015), scientific reasoning (General Medical Council 2009) and writing abilities (Goodman and Edwards 2014, McCurry and Chiavaroli 2013). This confirmed that the skills assessed by BMAT remain relevant to the contexts that the test is used for.

Relevant theoretical models were used to examine the thinking skills assessed by Section 1, the scientific reasoning skills assessed by Section 2, and the written communication targeted in Section 3. Theories of critical thinking and problem solving were used to present the cognitive processes assessed by Section 1 as abilities that can be developed, and to distinguish the test construct from models of intelligence (Black 2012, Fisher 1992). This exercise raised some interesting issues. In particular, we identified a need to explicitly define terms commonly used to describe the constructs assessed by admissions tests, and to situate BMAT in relation to these terms. Based on a review of literature from educational psychology and assessment (e.g. Kaplan and Saccuzzo 2012, Newton and Shaw 2012, Stemler 2012), key terms used in admissions testing were defined and applied to BMAT. As a result, the title of BMAT's Section 1 is currently being reviewed, to evaluate whether 'aptitude' is a suitable description of the abilities that are being assessed. Think-aloud studies conducted by Cambridge Assessment researchers on item types from Section 1 were also presented. This illustrated one of the ways that cognitive processes elicited by a test can be investigated, and also demonstrated how findings from research can inform the processes used in test design.

Theories of scientific problem solving (e.g. Dunbar and Fugelsang 2005) were used to consider the cognitive processes involved in answering Section 2 items, and to conceptualise them as searches in a problem space (Simon and Newell 1971). Linking Section 2 to theoretical perspectives on scientific reasoning identified complex interactions between subject-specific knowledge and more domain-general reasoning abilities (Klahr and Dunbar 1988, Zimmerman 2000), which are components acknowledged as important to consider during Cambridge Assessment's item authoring processes. However, it is recognised that further investigation of the balance between knowledge and novel problem solving could be beneficial for assessing scientific reasoning. This presents a possible avenue for further research that might be supported by technological advancements, which have been used to investigate scientific problem solving (Tsai, Hou, Lai, Liu and Yang 2011).

Consideration of the theories underpinning Section 3 was heavily informed by Shaw and Weir's (2007) work on examining writing. Section 3's Writing Task was investigated in terms of the cognitive processes that it aims to elicit. In particular, the discussion focused on knowledge transforming processes that are commonly assessed at higher levels of language proficiency (Scardamalia and Bereiter 1987). However, the retrospective review of example responses to BMAT Section 3 could be complemented with further research on the cognitive writing processes activated when responding to tasks. Key logging, eye tracking and verbal protocol analysis could potentially be used to investigate how candidates plan, organise and monitor whilst writing. The skills assessed by Section 3 are also regarded as examples of test takers' productive reasoning abilities, drawing on critical thinking and assessment research recommending that constructed responses are used to complement other formats commonly used in standardised testing (Butler 2012, Liu et al 2014).

The conceptualisation of BMAT sections as assessments of separate skills has also been investigated as part of cognitive validity. A key study confirming that it was valid to interpret Sections 1 and 2 as measures of two distinct skills was conducted by Emery and Khalid (2013a); this was presented to illustrate another method commonly used to investigate cognitive validity. Chapter 3 highlighted how important it is to consider the theory underlying an admissions test. It was argued that assessment providers have a responsibility to present theoretical reasons for assessing the cognitive processes targeted by examinations, and that theories should be investigated with research.

Context validity

BMAT's context validity was examined in Chapter 4, which stressed the relationship between context validity and cognitive validity. Designing tasks for

an admissions test such as BMAT requires careful consideration of various features, because the response format, test timing and task content can all influence the skills assessed by a test. These issues were considered in some detail when developing guidelines used to support context validity, so they are explored in some detail. Multiple-choice questions (MCQs) and tasks requiring constructed responses have specific advantages and disadvantages, so BMAT uses a combination of these task types across its sections.

The number of tasks to include in a test section is another feature of context validity that was considered for BMAT Sections 1 and 2, alongside evaluations of the time needed to complete typical test items. These considerations were informed by early research studies that investigated speededness in BMAT (Shannon 2005), that ultimately led to changes in the number of items included in BMAT sections. More recent studies monitoring time pressure in BMAT have been conducted by analysing omission rates, and an example of this was presented as a key study (Emery 2013a).

A number of threats to validity can be introduced or overlooked when constructing test tasks, and various steps are used to ensure that BMAT assesses the correct skills as intended. Cambridge Assessment's approach to authoring tasks uses detailed test specifications, review by subject matter experts (SMEs) and standardised processes to safeguard against threats to context validity.

A particularly important feature checked for all tasks is the knowledge required to successfully complete the task. Some tests mistakenly claim to include tasks that do not require any knowledge, when they actually mean that test tasks assume a certain level of non-specialist knowledge. For BMAT, tasks destined for Section 1 or Section 3 are checked against a threshold of everyday knowledge. Section 2, on the other hand, assesses the ability to apply subject-specific science and maths knowledge to novel questions. This makes it important to identify the aspects and level of subject knowledge that a test taker is expected to have when they take BMAT. A recent review of the curriculum underpinning BMAT Section 2 was conducted and described in the chapter, to illustrate how assessment experts can explicitly define a pool of assumed knowledge for a test. Once defined, the subject knowledge curriculum was used to support suitable test taker preparation. Furthermore, it allowed SMEs to check the science needed to answer an item correctly against the topics included in the curriculum. However, the checks relating to subject knowledge are not the only ones required to ensure context validity.

BMAT items are commissioned and stored in item banks in preparation for constructing test papers. Various SMEs are recruited to author, edit and vet items before they are placed in a BMAT item bank. A description of the multi-stage question paper production process was presented to outline how different SMEs review specific issues, first in items, and then in papers. The checks conducted during item commissioning, item editing, paper

construction and paper vetting were described, alongside rationales for their inclusion.

Another important part of context validity refers to the administration conditions associated with a test. For BMAT, it is critical that the test is administered securely and in a standardised way, so that test takers experience similar conditions when completing tasks. Cambridge Assessment's approach to administration uses strict test regulations and centre approval processes to monitor these issues. Furthermore, the advantages and disadvantages of various administration methods are continuously reviewed. For example, the possibility of using a computer-based (CB) testing model is regularly evaluated with consideration of the security and access issues associated with a change from paper-based (PB) testing. Although the discussion currently presented in Chapter 4 concludes that BMAT should continue to be administered in PB format, it is entirely possible that this will change at some point in the future. In terms of BMAT's context validity, this is one area that will undoubtedly require further research. In particular, the equivalence of completing CB and PB tasks will likely form the focus of future work on context validity.

Scoring validity

Chapter 5 focused on the processes used to minimise error and ensure that BMAT scores are meaningful. A range of statistical methods are used to safeguard BMAT from threats to scoring validity, and these are presented to provide the reader with an overview of operational validation processes that monitor BMAT sessions.

For the MCQ sections of BMAT, analysis is used to check that items are appropriately difficult and that they discriminate between test takers with low and high abilities. This ensures that the test is targeted to a suitable level for Sections 1 and 2. Rasch analysis is used to score these sections and report them. The approach to scoring taken by Cambridge Assessment produces a scale ranging from 1.0 to 9.0, where equal intervals in BMAT scale scores represent equal differences in candidate ability.

In addition to analysis that is used to monitor and produce scores, a number of analyses are conducted regularly on BMAT Sections 1 and 2. These show that BMAT sections have acceptable internal consistency and also indicate that items are free from bias in relation to gender and school type. The limitations of commonly used statistical coefficients are also relevant to discussions of internal consistency, so they are presented with some of the reasons that estimates of reliability are necessary, but not sufficient, indicators of test quality. Although the internal consistency coefficients of the sections could be improved, this might not be appropriate for BMAT due to the relatively multidimensional nature of the sections and the cognitive

validity arguments for designing the sections in this way. Interestingly, there are parallels between recent developments in admissions testing and shifts in language testing observed by Weir (2005) over a decade ago. An overview of these issues is used to contextualise the approach to scoring validity adopted by Cambridge Assessment Admissions Testing, and to distinguish it from a more psychometrically led approach that is prevalent in the US.

For the scoring validity of BMAT Section 3, the marking criteria and marker training procedures are crucial. These safeguard scoring validity by systematically monitoring and evaluating the subjective marks awarded by examiners. These are detailed in Chapter 5 alongside some of the statistical procedures used to review marker reliability. These marker standardisation and training procedures for BMAT Section 3 are informed by research from language testing contexts (Shaw and Weir 2007). However, there are opportunities for further investigation of this area, because the impact of training on Section 3 examiners has not been investigated directly.

It should be noted that the procedures used to evaluate BMAT's scoring validity are designed specifically for the context of the test's administration. Future changes to BMAT's administration may require greater focus on scoring validity. For example, BMAT's use in an increasing number of territories may require alternative scoring procedures to be considered. To date, groups of candidates have not been considered across BMAT sessions that occur at different points in a year, because these tend to take place in different locations and are accepted by different university departments. However, increasing globalisation and student mobility may necessitate scoring procedures that enable precise comparability of scores across sessions, most likely with statistical equating. These procedures sometimes require additional data to be collected, so developments will need careful consideration of logistical and security issues. Furthermore, Cambridge Assessment researchers may need to develop innovative methods of scoring to deal with use of BMAT in new contexts, and this represents a significant focus for development of the test.

Criterion validity

Investigating the relationships between test scores and other variables is a key consideration for assessments used in selection contexts, such as admissions tests. In particular, predictive validity is prioritised over many other aspects of validity when selecting applicants for job roles and university places. In medical selection, some researchers refer to correlations between on-course performance and test scores as 'the validity coefficient' (Cleland et al 2012:11), and predictive validity is emphasised over other forms when discussing admissions tests (e.g. McManus, Dewberry, Nicholson, Dowell et al 2013). In line with these established conventions, Cambridge Assessment has placed a historical emphasis on this aspect of validity. In particular,

our researchers have focused on BMAT's predictive validity and equity in published research (Emery and Bell 2009, Emery et al 2011). However, Cambridge Assessment Admissions Testing's contemporary approach to validity adopts the socio-cognitive framework (O'Sullivan and Weir 2011) and acknowledges that other aspects of validity are also relevant to admissions tests. This contrasts with the approach adopted by some other researchers, who treat predictive validity as the only form of validity that matters in selection contexts (e.g. Hopkins et al 1990).

In considerations of criterion-related validity, we heed Weir's (2005:13) warnings that 'no single validity can be considered superior to another. Deficit in any one raises questions as to the well-foundedness of any interpretation of test scores.' The tendency to primarily consider one type of validity over others has also been a concern for experts in the wider educational assessment community, who have reflected on some historical practices that prioritised particular forms of validity. For example, Newton and Shaw (2014) describe how conceptualisations of validity as the hypothetical agreement between test scores and a theoretical true proficiency led to an early focus on criterion validity. This developed almost accidentally, as researchers overlooked the limitation that operationalised criterion measures were flawed representations of true proficiency.

Therefore, Cambridge Assessment researchers consider a wide range of methodological and theoretical issues when planning predictive validity studies. A critical approach is required because various issues reduce the strength of relationships in selection contexts. Whilst corrections for attenuated coefficients are available (e.g. Sackett and Yang 2000), applying them uncritically in pursuit of a stronger 'validity coefficient' may not be appropriate in complex selection contexts. Indeed, corrected coefficients can hinder, rather than support, meaningful interpretation if applied without an understanding of common methodological challenges and how they might have impacted on the specific selection context of the study. In addition, concurrent validity in the context of BMAT was discussed to highlight that various admissions tests used for healthcare selection are assessing quite different constructs, rendering comparability studies unsuitable. Furthermore, there is little agreement on how potential for medical study should be conceptualised for an admissions test, so there is not an external framework suitable for benchmarking BMAT in concurrent validity studies. Development of a framework for selection to healthcare courses is a suitable area for medical educators to explore.

The authors of Chapter 6 present conceptual overviews of the theoretical issues and methodological challenges relevant to investigating criterion-related validity in selection contexts. Illustrative examples are used to introduce the issues to those who are unfamiliar with them, recognising that they tend to be exacerbated by common selection practices, and the impact of these procedures is easy to overlook.

Cambridge Assessment's approach to conducting and reporting predictive studies is also presented, in order to contextualise the research summarised in the chapter. This approach advocates reporting uncorrected coefficients alongside known information about the selection procedures used, which can be achieved by conducting situated studies in collaboration with admissions tutors. However, this is not presented as the only appropriate way of investigating criterion-related validity. Recently, Cambridge Assessment has been collaborating with the General Medical Council (GMC) to provide data for a UK Medical Education Database (UKMED). This initiative is described to illustrate how big data approaches can also contribute to understanding the relationships between test scores and other outcomes.

These developments present future research opportunities to investigate the criterion-related validity of admissions tests, particularly because inclusion of various other selection criteria could enable researchers to accurately describe the procedures that are used in practice. However, it must be recognised that large datasets do not eliminate the need to collaborate closely with admissions tutors and understand issues specific to their contexts. Therefore, the challenge for researchers conducting further work in this area is to embrace the opportunities provided by these developments, whilst remaining cautious in case of spurious findings that are not explained by theory. This can be achieved by guiding statistical analysis with *a priori* consideration of theory, and by complementing large-scale studies with smaller ones.

Consequential validity

In Chapter 7, the social impact of using BMAT was unpacked using the concept of consequential validity from the socio-cognitive framework. By applying a broad conceptualisation of consequential validity to the admissions testing context, McElwee, Fyfe and Grant extend arguments made in the field of language testing to the sphere of admissions testing, and also into medical education. In this regard, the social and ethical issues related to a test's use should be considered part of overall validity rather than as a separate element. Whether assessment experts refer to these issues as validity or not, it is generally agreed that they are important for the test developer to consider (Newton and Shaw 2014). In our view, omitting consequential validity from models of validity would allow test developers to argue that this aspect of assessment rests solely with test users, and this stance would be detrimental to educational assessment; therefore, consideration of social and ethical consequences should be integrated into models of validity. Integration with other aspects of validity is particularly important because the analysis of consequential validity presented in Chapter 7 showed that test use can impact on issues recognised as central to validity.

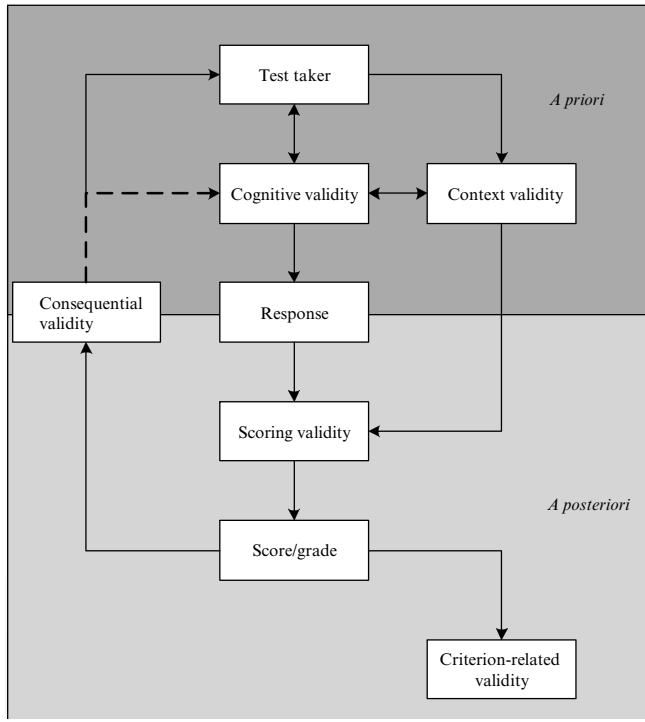
In previous applications of the socio-cognitive framework, much has

been stated about the symbiotic relationship between cognitive, context and scoring validity (e.g. Khalifa and Weir 2009) because these aspects constitute the core of construct validity. Revisions of the socio-cognitive framework have reflected this by explicitly referring to the two-way relationship between context and cognitive validity (O'Sullivan and Weir 2011). Similarly, this volume has emphasised how these elements interact for an admissions test; however, the admissions testing context presents an opportunity to identify other interactions that further extend the socio-cognitive framework, particularly in relation to consequential validity.

Methods used to select university applicants inevitably impact on widening access initiatives in higher education, which are important issues for policy-makers and society. In terms of access to the medical profession, the emphasis on widening participation is even stronger than in other disciplines. In a close examination of BMAT's consequential validity, the authors of Chapter 7 point out that consequential validity is considered not only as *a posteriori* to a test event as conceptualised in Weir's original framework, but also *a priori* due to the impact on following selection rounds and future cohorts of applicants. In this regard, it should be recognised that the test's impact on wider society can change the test taker population for further administrations of an admissions test. How the test is perceived can potentially change the applicant pool, which might have a knock-on effect for a professional workforce. Therefore, it is particularly important to investigate consequential validity and recognise this mechanism in the admissions testing context by revising the socio-cognitive framework (see Figure 8.1).

Cambridge Assessment has not ignored the consequential validity of admissions tests, and the key studies presented in the volume are evidence of that, but it is fair to say that this area has only been focused on relatively recently. This has partly been prompted by adoption of the socio-cognitive framework, but also because BMAT users have sought to understand how prospective medical students prepare their applications for medical school. This trend is reflected more widely in recent medical education research looking at selection, which has investigated how assessments are perceived by applicants and members of the medical profession (Cleland, French and Johnston 2011, Kelly, Gallagher, Dunne and Murphy 2014, Stevens, Kelly, Hennessy, Last, Dunne and O'Flynn 2014). Despite general worries that selection procedures might deter potential applicants from certain groups, these issues have not been viewed as aspects of validity. In medical education, consequential validity is only used to refer to issues that stem from test score interpretation (Downing 2003). We argue that conceptualising consequential validity in a broader sense would support the development of theoretical frameworks about the consequences of test use in medical education, where there have been calls for more theory when evaluating initiatives to widen access (Nicholson and Cleland 2015). This theory-based approach has

Figure 8.1 A revised socio-cognitive framework



been adopted by some medical education researchers looking at admissions; Niessen, Meijer and Tendeiro (2017) framed qualitative findings on the consequences of using selection methods as part of organisational justice theory. Survey research on test taker perceptions of selection methods was presented as a key study in Chapter 7 (Emery and McElwee 2014) and it may be useful to consider the results in light of wider social theories.

Selection to study medicine and dentistry is a key place where attention to theory can have an important impact. In addition to considering the technical and predictive components of selection methods, policy-makers should recognise that assessments at this stage potentially shape the attitudes and beliefs of future healthcare professionals (Röding and Nordenram 2005). In the Netherlands, research has compared the motivation and self-beliefs of medical students entering through competitive selection with those selected by lottery. Wouters et al (2016) found that the strength of motivation was higher in competitively selected students. Although these differences were not shown to be pervasive in the long term, they do warrant further investigation in other selection contexts. There is also evidence that the relationship

between selection procedures and motivation varies across studies, indicating that contextual factors could be important when investigating motivation in medical students (Wouters, Croiset, Schripsema, Cohen-Schotanus, Spaai, Hulsman and Kusurkar 2017).

Research from educational psychology may also present insights into these issues. Experimental work with children indicates that motivation and resilience are influenced by beliefs about the fixedness of their academic abilities (Dweck 2012, Yeager and Dweck 2012). Whilst it would be a mistake to apply these ideas uncritically to adolescents applying for university study, we should consider the self-beliefs promoted by selection procedures, and whether their impact might differ on the subgroups present in applicant pools. Consequential validity poses specific questions about how the constructs we assess can influence those being assessed. Answering these theoretical questions can potentially inform the ways that universities communicate about selection to prospective applicants.

Researchers should investigate how assessment constructs are perceived, not just by university stakeholders, but also by test takers. Cambridge Assessment's approach to admissions testing recognises that scores on all such tests, even those grounded in the psychometric approach to intelligence, are 'a function of innate talent, learned knowledge and skills, and environmental factors that influence knowledge and skill acquisition' (Kuncel and Hezlett 2010:339). Therefore, Cambridge Assessment's admissions tests, which are constructed with a focus on skills that can be developed, should not be conceptualised purely as measures of innate attributes. This has been communicated to admissions tutors and other assessment experts; however, we do not fully understand how test takers perceive tests such as BMAT and, importantly, how they understand their performances on them. Despite BMAT's explicit focus on skills that can be developed, do admissions tests encourage biomedical trainees to believe they were born smart enough to become a doctor or dentist, and that other people were not? If so, what is the impact of this, if any, on their learning and their future clinical practice? Perhaps even more crucially, what impact is there on test takers who come to believe they were not born with the genetic endowment to become a doctor? Furthermore, these considerations must inform the current search for evidence-based 'non-cognitive' criteria (Hecker and Norman 2017). Bearing in mind that tutors will need to communicate decisions to those who are ultimately unsuccessful at entering the healthcare professions, what does it mean to not have the integrity for entering medical study? Understanding these issues can potentially develop theories about student motivation and also inform higher education policy.

The reflections on consequential validity presented in this volume, and particularly in Chapter 7, are initial steps towards addressing this aspect of validity in admissions tests. There are many directions and areas of

investigation that stem from the questions posed by consequential validity. One example is represented by the dashed line (to indicate a tentative relationship for investigation) linking consequential validity to cognitive validity in Figure 8.1. In the admissions context, if the selection policy of a university treats an assessment in a way that is incompatible with the targeted construct, then the meaning of the score can potentially be changed. Consider a university's policy on accepting results from examinations that have been sat more than once. If the assessment targets an ability that is beneficial for a particular field of study, the rationale for using the selection method is normally that the ability is associated with study success, either incrementally or to a pre-requisite level. In this situation, previously achieved scores are indicators of ability from earlier in the developmental process; they are not relevant to decision making at the point of application, and the selector should accept results from the most recent sitting of the exam.

One example that illustrates how this issue manifests in practice is when universities decide whether to accept A Level grades achieved in resits. If the grade at first attempt is the only one considered, this changes the nature of the construct that the score represents. The A Level cannot be conceptualised as mastery of a knowledge-based curriculum in this situation, because the policy dictates that the first attempt stands. Mastery of a knowledge-based curriculum can theoretically be improved upon and developed, but the policy has instead changed the meaning of the A Level grade that is accepted. Of course, A Level grades at first attempt are influenced by many different factors and universities may have good reasons for treating them in this way. McManus et al (2005) observe that A Levels could be indirect indicators of motivation or commitment, and conceptualising them in this way may be predicated on the applicant studying multiple subjects at the same time. However, universities should consider the theoretical reasons for using an assessment outcome in a particular way.

These issues also apply to assessments that claim to assess innate abilities. For these measures, as the trait being assessed is theoretically fixed, test scores should not vary across multiple attempts. In fact, multiple test attempts can be conceptualised as parallel evaluations of the same innate trait, and the most valid score to consider would be some kind of average across the attempts. Decisions about accepting resits are often made due to practical concerns about the number of applications that a university can consider in a cycle. Biomedical courses sometimes provide empirical reasons for not recognising A Levels that have been re-examined, using data to show poorer outcomes for students admitted with resits. However, policy-makers should also attempt to understand the mechanisms that drive these outcomes. The idea that consequential validity can influence cognitive validity highlights the need to reflect on the ways that commonly used selection criteria are conceptualised. The interactions between consequential validity, other aspects of

validity and wider social theory represent areas to be explored with future research.

Conclusion

In Chapter 1, Saville proposed that the socio-cognitive framework developed in language testing could guide comprehensive evaluation of BMAT's validity. This volume has used Weir's (2005) socio-cognitive framework to present key aspects of test validity, and demonstrated how they can be used to consider validity of an assessment used in selection for medical study. Application of the socio-cognitive framework to BMAT demonstrates its flexibility as a model for test evaluation, and provides an example of how it can be used to focus attention on aspects of validity, in an assessment other than a language test. Some aspects of validity identified in the socio-cognitive framework are commonly overlooked in the admissions testing context, despite being considered regularly by researchers working in language testing. However, none of the issues covered can be considered trivial and each chapter successfully argues that the aspect of validity focused upon is important. By considering each aspect in turn, we have shown how they relate to the ways BMAT was developed, how it is currently administered, and how its validity is continuously monitored.

Throughout the volume, we have reiterated that the separate chapters of the book do not represent isolated issues relating to the use of BMAT. Rather, the chapters, and the socio-cognitive framework itself, provide a structure for systematic investigation of validity as a unitary concept. Nevertheless, organising the issues in this way can give the mistaken impression that they are discrete topics. Therefore, it is important to reiterate that the aspects of validity described throughout this volume are interconnected. This volume demonstrates that Cambridge Assessment's approach to admissions testing fits particularly well with a socio-cognitive framework that conceptualises validity as unitary. Various aspects of BMAT's validity are considered necessary but not sufficient to ensure that inferences based on test scores are valid. In this approach, validity is conceptualised on a continuum, but test quality is not linked simply to isolated coefficients representing psychometric quality. Evidence that each aspect of validity has been considered for BMAT contributes cumulatively to the confidence associated with use of test scores. This dissuades test developers from focusing blindly on one or two aspects of validity at the cost of others, which has been a historical issue in educational assessment, as demonstrated by a quote from the 1966 edition of the *Standards*: 'Too frequently in educational measurement attention is restricted to criterion-related validity' (1966:6).

This collection of chapters is not intended to be an exhaustive compilation of research on BMAT, but rather to give an insight into some of the ways

that the test has been evaluated. Hopefully, readers from various disciplines will have found the description and discussion of Cambridge Assessment's approaches useful. This volume has demonstrated how a multidisciplinary approach spanning language testing and admissions testing can be beneficial. It would be good if sharing this work with medical educators, language testing researchers and admissions test developers could encourage collaboration across subgroups of educational assessment experts, to share expertise and best practice in a way that benefits various forms of assessment.

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.