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

4 Building fairness and appropriacy into testing contexts: Tasks and administrations

Mark Shannon
Cambridge Assessment Admissions Testing
Paul Crump
Cambridge Assessment Admissions Testing
Juliet Wilson
Cambridge English Language Assessment

4.1 Introduction

In Chapter 3 of this volume, Cheung and McElwee focus on the theoretical basis for the cognitive processes assessed by BMAT sections. They point out that designing test tasks is necessarily intertwined with the cognitive processes targeted by items. The aim of the present chapter is to closely examine context validity, which includes the task design considerations that can influence whether BMAT tasks are assessing what they are intended to measure. Similarly, features of the test administration can also impact candidates' cognitive processes and threaten the validity of an examination. Context validity is concerned with the conditions under which a test is taken. It asks whether, and to what extent, the characteristics of the test tasks and their administration are fair and appropriate for candidates (Weir 2005). Principles of fairness dictate that all candidates should have the same experience wherever in the world they take a test.

As mentioned in the previous chapter, context validity exists in a close relationship with cognitive validity, in that it includes the representativeness and authenticity of the test tasks to the wider domain. Task design decisions regarding the response format, method of marking and number of tasks in a section also impact on the ways that a test can be scored, so context validity affects scoring validity.

Features of the task can impact on the testing situation in many ways. For example, the length of time allowed to complete a task or tasks must consider the impact on the cognitive processing of candidates, particularly

if there is not sufficient time to complete the items. A key research study on time pressure is presented in this chapter to illustrate how this issue can be investigated, and how analysis of test data can inform the quality assurance procedures used in paper production. These procedures include a range of checks for each section that ensure tasks elicit the cognitive processes outlined in Chapter 3.

Context validity also encompasses questions concerning test administration conditions. Some of these impact on the security and uniformity of testing conditions, which are key issues when considering high-stakes exams. The present chapter describes how Cambridge Assessment deals with this wide range of context validity issues, both in terms of research and operational practice. In part 4.3, the key considerations in task design are outlined, along with the checking procedures that are used to ensure design decisions are maintained in practice. Following this, Cambridge Assessment's approach to standardising administration conditions is presented, along with examples of the inspection process that is used to monitor test centres. Firstly, in the next part of this chapter, we examine context validity as outlined in the socio-cognitive framework (O'Sullivan and Weir 2011, Weir 2005) and situate it in relation to BMAT.

4.2 Context validity and BMAT

Aspects of context validity are generally classified as features of the task or features of the administration conditions. Under the task aspect come considerations such as the authenticity of the types of tasks, response format and rubric. Also included are considerations that apply to an entire test or test sections, which can include multiple tasks. Examples of these are the time constraints for completion, the weighting and order of items, and candidates' knowledge of the marking criteria. For Sections 1 and 2 of BMAT, these issues are monitored in the item authoring and paper production procedures used to construct versions of the test. Writing tasks for Section 3 undergo similarly rigorous production procedures, but equally important are the rubrics and processes that ensure valid assigning of marks, which are discussed in Chapter 5. The 'administration conditions' aspect includes a consideration of the uniformity and the security of the testing conditions. The logistics of ensuring security should not be underestimated for large-scale examinations, and BMAT's increasing use internationally presents challenges to maintaining standardised test administrations.

Any of these factors, unrelated to the candidate's ability on the construct of interest, could impact test performance. In ensuring test validity it is essential that the test provider understands the effects of such features on performance and ensures that they are controlled and standardised as far as practically possible, both between test papers and between testing situations.

The context validity component within the socio-cognitive validation framework can be used to pose specific questions as follows: Is there any evidence that the response format is likely to affect performance? Are the marking criteria explicit for the candidates and the markers? Is the timing of each part appropriate? Is the content knowledge suitable and unbiased? Are the administration conditions satisfactorily consistent and secure?

The representativeness, appropriateness and authenticity of the tasks in a test are what give us faith in the generalisability of test results to what we are trying to measure. Response format (e.g. multiple-choice questions (MCQs) versus constructed response) is often constrained by considerations such as scoring validity (e.g. the desire to have items that are marked objectively) and practicality (e.g. the speed and lower cost of marking MCQ items). These issues are considered carefully in language testing, where it is common to use a mixture of response formats across the four skills commonly evaluated in a test (Elliott and Wilson 2013, Galaczi and ffrench 2011, Khalifa and Weir 2009, Shaw and Weir 2007). It is also considered good practice to use more than a single response format in a test when assessing higher-order reasoning (Liu et al 2014), as each response format has its advantages and disadvantages. The timing of the test is an important consideration but one that is also often constrained by practicalities. Speededness[1] may be a part of the test construct but the time pressure should not be such that candidates are unable to complete the test within the time allocated or are unduly stressed. Candidates should be made aware of the timing, number of items, weightings, marking criteria and any penalties for incorrect responses. The task rubric must be explicit, unambiguous, simple and brief yet comprehensive. No candidate should be able to misinterpret the test tasks.

A crucial threat to the context validity of a test (and the reputation of the test provider) is the potential for malpractice on the part of the candidate or the test centre (Cizek 1999). The higher the stakes of the test, the more of an issue cheating is likely to be. For this reason, the security of administration conditions is a vital concern of both the test provider and stakeholders, including test takers themselves who must perceive the test as fair. The increasingly sophisticated technology available for cheating in examination conditions means that detection (post-test), as well as prevention, is a responsibility of the test provider. Admissions tests for biomedical and dentistry study are certainly high stakes and a summary of the statistical approaches used with BMAT for malpractice detection is available in Chapter 5 of this volume. For those interested in the more technical aspects of statistical malpractice detection, our approach in this area is informed by work

1 Speededness refers to 'the situation where the time limits on tests do not allow substantial numbers of examinees to *fully* consider all test items' (Lu and Sireci 2007:29, emphasis added). In contrast, a 'power test' is one where the correctness of the answers is key, regardless of how long test completion takes.

on Cambridge English exams, as outlined by Bell (2015) and discussed by Geranpayeh (2013). The present chapter focuses on the standardised procedures and security checks used by centres administering BMAT, and the centre inspections used to ensure that the test is administered according to Cambridge Assessment's standards.

The following part of this chapter, part 4.3, addresses the aspects of context validity that focus on BMAT tasks. This includes a case study of work conducted to revise and define the content knowledge examined in Section 2 and a key research study into the appropriateness of the time constraints in BMAT.

4.3 Cambridge Assessment practice: Task features

Response format and task design

Two types of response format are used in BMAT and this is a fixed feature of the test, as is the number of items per section. Sections 1 and 2 of BMAT are multiple-choice format (with each item weighted equally) whereas Section 3, the Writing Task, requires candidates to construct a brief essay response to a structured prompt. These two response types, both of which are likely to be encountered in the future course examinations of successful applicants, can be seen as representing each end of the response-type continuum. There are advantages and disadvantages to both formats. Here we discuss the context validity considerations related to each response format.

Multiple-choice questions/items

MCQs are a popular form of standardised assessment because they are objective, low cost and it is possible to mark them quickly after the test session. Liu et al (2014) point out that MCQs typically cost more in assessment development than constructed-response items, but are cheaper overall due to the cost of marking constructed-response tasks. This observation applies to BMAT MCQ items, which go through multiple stages of checks in the question paper production process (these are outlined later in the chapter under test content). Responses to these are then objectively marked using optical scanners. Compared with constructed-response tasks, MCQs fit better with psychometric models used to investigate internal consistency and reliability, because a greater number of MCQs can be included in a test with limited time (see Chapter 5 for details). Also, quality assurance processes can be automated to evaluate test sections based on the responses in a session and this post-test evidence of validity is available to the test providers before results are released. Given that medical schools using BMAT work to tight timescales when making selection decisions, these are substantial advantages

over constructed-response items, which take longer to mark reliably and quality assure. However, there are some criticisms of MCQs that need to be considered by test providers. Chiefly among these are observations that the reasoning used to answer an MCQ is different from the reasoning employed in non-test settings, because test takers rarely select from a defined set of options. This is often raised in relation to listening and reading exams (Field 2013, Khalifa and Weir 2009); for example, Field points out that MCQs in listening exams often require candidates to engage processes that fall outside of a real-world listening event, such as disconfirming available response options.

Similarly, answering MCQs can require BMAT candidates to engage reasoning that is somewhat different from the reasoning involved in clinical practice. Indeed, Sam, Hameed, Harris and Meeran (2016) observe that clinical medicine is often nuanced, which runs counter to the idea of a single correct answer as assumed by MCQ formats. However, it should be noted that BMAT targets the potential for biomedical study rather than practice in a clinical environment, and MCQs are used as an assessment tool in undergraduate studies. Their use is an established method in medical education contexts (Downing 2002), where MCQs are used to evaluate both factual recall and higher-order cognition (Palmer and Devitt 2007).

Furthermore, there is evidence that constructed-response items correlate positively with MCQs (Klein, Liu, Sconing, Bolus, Bridgeman, Kugelmass and Steedle 2009, Rodriguez 2003), indicating that MCQs can be valid assessment tools when constructed appropriately. In reference to the medical education setting, Downing (2002:240) points out that in order to produce valid MCQs, 'item writers must have the willingness to invest considerable time and effort into creating effective MCQs'. Cambridge Assessment invests a great deal of time not only authoring items, but also in reviewing, editing and vetting them. A process-driven approach is used to review items and consider the plausibility of the incorrect response options (distractors), the cognitive processes needed to reach the correct answer and the number of response options available to candidates.

Another issue to consider is whether particular formats might advantage or disadvantage particular groups of test takers, and this question has been raised in relation to MCQs, particularly in terms of gender differences. However, evidence of gender bias in MCQs is mixed. Large studies and meta-analyses conducted by Arthur and Everaert (2012), Buck et al (2002) and Du Plessis and Du Plessis (2009) did not find any systematic bias in MCQs. Similarly, a study conducted by Cambridge Assessment researchers on General Certificate of Secondary Education (GCSE) data did not show that MCQs advantage a particular group over others (Bramley, Vidal Rodeiro and Vitello 2015). However, a clear bias exists in MCQs which penalise test takers for incorrect responses, as this is linked with differential response rates

between males and females. Baldiga (2014), Kelly and Dennick (2009) and Hirschfeld, Moore and Brown (1995) found that a male advantage exists for MCQs which use negative marking in diverse disciplines including history, medicine and accounting. Baldiga (2014) argues that it is the high-stakes nature of the environment, coupled with score-awarding that exacerbates a socialised (rather than cognitive) difference between males and females. Therefore, negative marking should not be encouraged in high-stakes testing and BMAT does not penalise incorrect responses in MCQ sections of the test.

Due to their efficiency, reliability and objectivity, the majority of testing time in a BMAT session is allocated to MCQs in the form of Section 1 (60 minutes) and Section 2 (30 minutes). However, an essay task (for which 30 minutes is allowed) is also used to assess productive reasoning in Section 3, which requires a constructed response.

Constructed-response essay task

Developing and producing a written argument is a key skill for higher education study that is not possible to assess with MCQs. This provides a strong theoretical rationale for including a constructed-response test section that complements the MCQ sections of BMAT. For a discussion of these theoretical issues, see the cognitive validity arguments outlined by Cheung and McElwee (this volume).

The essay task for Section 3 was originally modelled on the US Medical College Admission Test (MCAT) in use at the time, and in the early years of BMAT the structure and wording from the original Oxford Medical Admissions Test (OMAT) was followed. This has been modified in order to make the rubric clearer and more accessible. For example, early questions asked candidates to produce a 'unified essay', meaning a structured and coherent argument rather than unconnected statements. This phrasing was discontinued in case it should prove confusing or unfamiliar for candidates. In other words, the instructions were improved to better elicit the targeted cognitive processes from candidates.

The BMAT Writing Task prompts are highly structured in order to guide test takers through the task, ensuring that even weaker candidates are supported to produce a script that can be scored. On each Section 3 paper, candidates are presented with three questions, from which they must choose one; in broad terms, the three questions will cover a general, a scientific and a medical topic. Topics are carefully chosen and the questions are vetted by two independent consultants to ensure that they are accessible to a diverse international candidature. Each Writing Task question presents a statement and asks the candidate to explain what it means. The candidate is then asked to argue to the contrary and finally to summarise or conclude with reference to the wider context of the statement. An example question is available in Box 4.1.

Box 4.1 Sample question from BMAT Writing Task

When treating an individual patient, a physician must also think of the wider society.
Explain the reasoning behind this statement. Argue that a doctor should only consider the individual that he or she is treating at the time. With respect to medical treatment, to what extent can a patient's interests differ from those of the wider population?

The prompt used in BMAT Section 3 encourages the use of knowledge-transforming strategies and processes (Scardamalia and Bereiter 1987), which produce more advanced writing, by providing an explicit argument for the test taker to conceptualise as a rhetorical goal. This encourages writers to form mental representations of their main points, which is essential to producing a cogent argument. Shaw and Weir (2007) argue that a writing task should be designed to elicit a response with a clear purpose and the prompt should make this explicit to the test taker. According to Weigle's (2002) categorisation of written discourse, there are six purposes, or dominant intentions, that can be specified for a piece of writing (Box 4.2).

Box 4.2 Categories of dominant intention from Weigle (2002:10)

Metalingual mathetic (intended to learn)
Referential (intended to inform)
Conative (intended to persuade or convince)
Emotive (intended to convey feelings or emotions)
Poetic (intended to entertain)
Phatic (intended to keep in touch)

By focusing BMAT's Section 3 prompt on argument, it is made clear that the dominant intention of the written response should be conative (intended to persuade or convince). In addition to specifying an argument, the structured prompt provides questions that should be addressed as part of the written response. Answering these questions requires a candidate to generate ideas on the topic area, as source material is not provided for candidates to reorganise or reproduce. Instead candidates are expected to draw on relevant general knowledge and develop ideas from these to construct an argument.

The answer sheet provided to candidates is also designed to encourage planning before writing a structured argument. Only one side of A4 is provided for the actual essay and candidates are told that no additional answer sheets may be used. Test takers with permission to use a word processor

are instructed not to exceed 550 words. This limitation on the length of the response means that candidates need to plan and structure an essay that will fit in the space available. Whilst a single side of A4 is enough to produce an example of extended writing, it is also intended to be easily manageable for BMAT candidates in the 30 minutes provided for Section 3, even when accounting for the time needed to select a question; therefore, there should be time that is allocated to planning, and space is provided for this on the question paper.

The design considerations required for MCQ and constructed-response tasks have been discussed here, focusing on matters relevant to the tasks included in BMAT. These issues are monitored in early stages of the question paper production process, which focus on evaluating items and tasks in isolation. Also monitored are issues that apply across entire test sections, such as the time allocated to complete all of the items in Section 1 or Section 2.

Test timing

Speededness is a feature of BMAT and a part of its test construct because test takers are expected to engage reasoning processes efficiently to complete questions. However, it is important that the time pressure is not excessive, so that the majority of test takers attempt every item, particularly for the MCQ sections.

In the first year of BMAT, 2003, there were slightly higher numbers of items in the two MCQ sections than in the years that followed: 40 items in Section 1 (Aptitude and Skills) and 30 items in Section 2 (Scientific Knowledge and Applications). The time allowance was the same as in the current test: 60 minutes for Section 1 and 30 minutes for Section 2. Due to finding higher than expected omit rates in the 2003 test, the numbers of items were reduced (for the 2004 test) to 35 for Section 1 and to 27 for Section 2.

Shannon (2005) conducted statistical investigation of the BMAT 2004 test items and again found potential evidence of excessive time pressure for Section 2, with omit rates rising towards the end of the section. As a result of these findings, the number of test items was not reduced any further but the recommendation was made to limit the number of time-consuming or complex items (e.g. those requiring candidates to answer a number of parts in order to gain a single mark). The number of BMAT items has therefore remained at 35 (in 60 minutes) for Section 1 and at 27 (in 30 minutes) for Section 2 since 2004.

These studies also informed guidelines for authoring items that consider the length of time required to read a question fully, carry out necessary calculations or apply reasoning. More recent studies have been used to monitor the time pressure of BMAT items and investigate hypothesised group differences about their impact. An example of this work is presented below as a key study.

Key study: Are the time constraints of BMAT appropriate? (Emery 2013a)

Introduction

One aspect of context validity is whether the time limits of a test are appropriate or overly pressured. Here we summarise aspects of a study investigating this issue in BMAT (Emery 2013a). Time pressure is an intended feature of BMAT but the time pressure of a test should not be such that the bulk of candidates are unable to finish the items or be forced to guess response options by the time pressure. Each of the three BMAT sections has its own, separate time allowance and all items in the two MCQ sections (Sections 1 and 2) are scored equally. Items in BMAT Sections 1 and 2 are intended to increase in difficulty throughout the paper (based on the judgement of the item writers). Items of each subtype (e.g. biology, chemistry, physics, mathematics) are interspersed throughout Section 2, but within each item subtype, the items judged to be easier are positioned earlier in the paper. An upward trend in guessing is therefore expected with item position in the paper.

Omit rates (the proportion of candidates that do not respond to an item) in excess of around 5% of candidates may be a cause for concern in non-MCQ examinations (Elliot and Johnson 2005), possibly indicating an unclear or difficult question. Given that BMAT is MCQ and it is advisable for candidates to guess items they do not know, omit rates for BMAT items are expected to be very low, and high omission of items might be suggestive of excessive time pressure.

The appropriateness of the BMAT time constraints was investigated using item-level response data from 2010 to 2012. The study also investigated whether the impact of the adverse effects of time pressure differed by gender. This is of particular interest given observations that male candidates have tended to score slightly higher on the MCQ sections of BMAT, as discussed in Chapter 2. The analysis for this study assumes that candidates work through the test in the order the items are presented in the paper, therefore if time pressure is excessive one would expect higher omit rates at the end of sections. The summary here focuses on omit rates, although the full report by Emery (2013a) also looked at other statistical indicators that may indicate guessing, such as item facility, item difficulty and item fit. As these largely confirmed the findings from analysis of the omit rates, these additional statistics are not discussed in the present summary; however, descriptions of these statistics and their application in test validation are available in Chapter 5.

Research questions

Is there evidence of excessive time pressure in BMAT Sections 1 and 2? Is there any evidence that females are affected more adversely by the time constraints in the test?

Data collection

Six item-level response datasets from BMAT Sections 1 and 2 were analysed (test years 2010, 2011 and 2012, whole cohort data). Candidate gender was captured at the time of test registration. All items were in MCQ format, apart from a single item that required a numerical response in BMAT 2010. Since 2011, all Section 1 and 2 items have been in MCQ format.

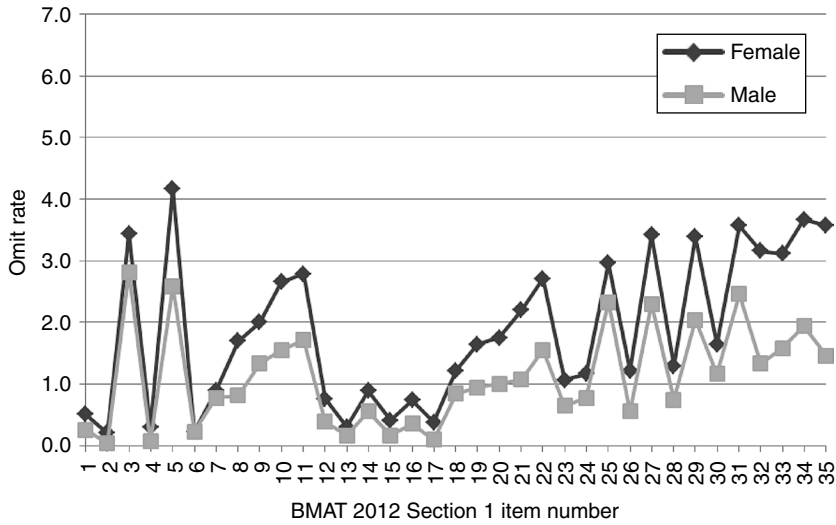
Datasets contained candidate gender and responses to each of the items (A/B/C/ . . . or 'omitted'). Candidate numbers in each cohort were as follows:

- BMAT 2010 N = 6,225 (57% female)
- BMAT 2011 N = 6,230 (57% female)
- BMAT 2012 N = 7,046 (56% female).

Results

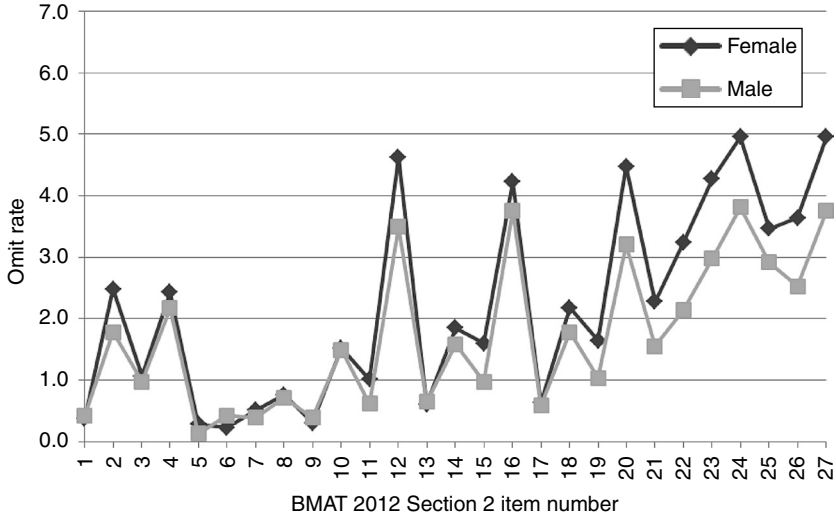
Figure 4.1 and Figure 4.2 plot the omit rates for BMAT Section 1 and 2 by gender, for the 2012 administration. Charts for all years were originally reported by Emery (2013a) and the results showed a similar pattern in each, with the largest difference in 2012.

Figure 4.1 Omit rates for BMAT Section 1 items (2012)



As shown in Figure 4.1 and Figure 4.2, a slight increase in omit rates specifically towards the ends of BMAT Sections 1 and 2 was evident for all three of these test years. Omit rates in all three years were low, however, with the values for items at the ends of the test sections amounting to less than 5% of the candidates failing to respond. There was also a trend for a higher

Figure 4.2 Omit rates for BMAT Section 2 items (2012)



proportion of females than males to omit items towards the end of the sections, and this was most markedly the case in Section 1 of BMAT in the 2011 and 2012 administrations. However, it is important to note that the gender differences in omit rates are very small; in cases with the largest differences this was 4–5% of females versus 2–3% of males. Other statistics indicated similar item fit for the later items in a section when compared to earlier items in a section. This suggests that the candidates who filled in a response for these end-of-section questions (i.e. the vast majority of candidates) were not guessing disproportionately at these compared to earlier items.

Discussion

The time pressure in BMAT was not excessive for most candidates based on the statistical evidence in this study. Both the Classical and Rasch item statistics did not suggest an unexpected decrease in candidate performance for items specifically towards the end of the test sections that would indicate their running out of time. Nor was there evidence that female candidates performed worse than male candidates on items towards the ends of the test sections. This latter finding is supported by Differential Item Functioning (DIF) analyses of BMAT items by gender (Emery and Khalid 2013a; see Chapter 5 for an outline of this study).

Omit rates *did* appear to show a clear increase towards the end of both test sections and this was more apparent for the female candidates. However, even in the years with the greatest number of omissions, the omit rate for females was only around 4–5% of candidates (compared to around 2–3%

of the male candidates). This indicates that relatively few candidates were unable to complete their responses within the time allowance, albeit a very slightly higher proportion of females than males. The slight difference in omit rates for males and females towards the test section ends did not translate into differences in the average number of correct responses for males and females on these items. It therefore seems unlikely that time pressure effects could explain the slightly lower performance of females on Sections 1 and 2 of BMAT overall in these years. This finding reflects that of Ben-Shakhar and Sinai (1991), who found a consistent pattern of greater omission rates among females in a battery of aptitude and selection tests but concluded that gender differences in guessing tendencies account for only a small fraction of the observed gender differences in multiple-choice tests.

The obvious caveat of this study is that candidates may answer items within BMAT Section 1 and within BMAT Section 2 in any order they wish because the test is paper-based (PB). The assumption was made for the purpose of these analyses that candidates tend to work through items in the order they appear in the test. It is possible, though, that the adverse effects of excessive time pressure could be manifest *throughout* the test sections rather than affecting only those items towards their ends, resulting in increased guessing (i.e. higher item difficulty and lower discrimination) throughout the test than might be obtained with a greater time allowance. Indeed it could be argued that the existence of omit rates at all is suggestive of time pressure given that there is no penalty for guessing.

Further research may therefore be warranted on the timing of BMAT. Manipulating the time allowance in experimental participants to assess its impact on item functioning is one potential method. Observational, interview and questionnaire data from live BMAT candidates would provide a valuable source of evidence. It will be particularly important to investigate the timing of BMAT should any changes be made to the test in future years. Omission of item responses seen in the BMAT data (albeit very minor) is a peculiarity given that there are no penalties for incorrect guessing. Cambridge Assessment is currently enhancing the free resources and support for BMAT preparation available on its website, and emphasising to candidates the advantage of attempting all questions is key. These observations also apply to Section 3, where the timing of the test has not been investigated as systematically as it has for the MCQ sections. Although responses submitted for the Writing Task are generally similar in length, it might be the case that time pressure has an impact on how candidates engage the cognitive processes involved in writing.

This study highlights the importance of considering the time available to complete a test as part of context validity and the value of investigating this issue with research. Earlier studies resulted in changes to the number of items included in BMAT sections, whereas the findings of this study confirm

the suitability of BMAT's format, and inform the processes that check the appropriateness of test items.

Test content – knowledge, suitability and freedom from bias

Ensuring that the content of BMAT is of a suitable level of difficulty targeted to the intended test takers is a central aspect of context validity. Another issue relevant to context validity is the knowledge related to answering items and the topic areas associated with test tasks. For test sections that do not specifically include bodies of knowledge (Sections 1 and 3), task content must be checked carefully to confirm that the context used to present an argument or problem would be encountered in everyday settings. They are also reviewed to ensure that answering an MCQ correctly or composing a suitable written argument does not rely on subject-specific knowledge.

Across Cambridge Assessment, guidelines are used to ensure that exam questions do not include emotive topics that can influence the performance of candidates, or particular subsets of test takers. These guidelines are used in the design of admissions tests; however, it is acknowledged that topics used in BMAT Section 3 might include more sensitive issues than normally considered acceptable throughout Cambridge Assessment, in order to authentically represent the issues that biomedical students will need to consider in their studies. Due to this relaxation of the guidelines, all Section 3 questions are scrutinised carefully by an assessment manager to evaluate potential for bias against specific groups. Topics that might evoke a different emotional response from subsets of candidates are avoided, even if they would be encountered in medical study. For example, medical issues more likely to seriously affect one sex over another, such as fertility or abortion, are avoided. Similarly, content for all sections avoids referring to religious or ethnic issues in their context.

An appropriate coverage of topics should be maintained for test sections that include subject-specific knowledge, such as Section 2. Because candidates are expected to be familiar with scientific knowledge and apply it to novel problems, it is important to define the scope of the topics that might be included in BMAT Section 2. To illustrate Cambridge Assessment's approach to specifying the science knowledge that underpins Section 2, a case study of a recent specification revision is presented in the next part.

Case study – Revising the BMAT Section 2 knowledge specification

In 2014 a revision of BMAT Section 2 Scientific Knowledge and Applications was undertaken by Cambridge Assessment Admissions Testing, with the aim of updating the syllabus and maintaining its relevance for biomedical education. This case study details the circumstances that prompted the revision

and the intended outcomes, as well as the work carried out to shape and develop the new Section 2 specification.

An important consideration in BMAT's development was that preparation should not require students to invest large amounts of time or money and should complement a student's school study. As BMAT is typically taken early in students' final year at school it was decided that the specification should cover topics that students would have been expected to study up to the age of 16 by the end of their GCSE examinations in the UK (approximately 18 months prior to BMAT).

When BMAT was developed it was used initially by universities in England, and therefore the National Curriculum for England, Wales and Northern Ireland, which outlined the compulsory science curriculum in state schools up to age 16, was an appropriate basis for the test content of Section 2. The intended message to test takers was that the core knowledge required for BMAT was already familiar to them through their compulsory schooling and significant amounts of new learning should not be required; it should instead be a matter of revision to refresh their understanding. At the time, the National Curriculum specified in detail the content to be covered, and this was reflected in the GCSE specifications from the major UK examination boards and textbooks, so there was plenty of information available to students preparing for the test, including those from overseas. On this basis, the content specification for BMAT itself was relatively brief, giving a broad overview of the test and topics that might be examined but referring students to the National Curriculum documents for further information.

Changes to the National Curriculum resulted in a less detailed programme of study and an increased diversification of curricular pathways to achieving GCSEs in Science and Mathematics. The changes made the task of ensuring that BMAT Section 2 contained only content covered in state schools by age 16 more difficult. Therefore, it was decided that a review of BMAT Section 2 content and the creation of a more detailed test specification was necessary to ensure candidates were supported, which comprised three main phases:

1. Compilation of draft specification.
2. Consultation with university stakeholders.
3. Trial by BMAT item writers and international experts.

Compilation of the draft specification

The first stage in establishing the basis for the new content specification was to conduct a review of major GCSE double-award Science and Mathematics syllabuses across five major UK examination boards to establish the breadth of topics that were encountered by potential BMAT test takers.

Senior examiners evaluated the specifications to identify the areas and

specific topics that were common across several examination boards, to derive the core material that would form the revised BMAT curriculum. Examiners identified the general topic areas, the details of the sub-topics that were covered in common, and commented on the overlaps in specifications (e.g. the depth of knowledge expected on the topic, diversity of exemplars used by the exam boards, and which boards did not include the topic).

The examiners were permitted to recommend the inclusion of some additional topics, provided they were accessible and easy for able students to learn independently. In these cases, one or more of the following justifications were required:

- topics judged to be essential to a core understanding of the particular science, even if there was less commonality in their appearance on the various exam board specifications
- relevant scientific principles taught earlier than GCSE that should be included for completeness (for students' reference)
- details or examples that draw links between topics to promote understanding of the inter-relatedness of science as a discipline
- topics of particular relevance to the study of medical/biomedical sciences.

Consultation with university stakeholders

A round-table discussion with senior academics involved in student selection and medical or biomedical education at the universities using BMAT was organised to refine the draft specification. Thirty-nine broad topic areas had been distilled from the examiners' initial review of GCSE curricula, comprising over 500 sub-topics. Each topic and associated sub-topics were discussed in turn and three questions were used to guide the discussion: Are the knowledge and concepts important for biomedical study? Is the context provided for the underlying principles relevant? Is the type of thinking that a topic affords important or useful in studying medicine, dentistry and veterinary medicine?

The first two questions are central to the construct validity of BMAT Section 2 and assuring its relevance for students who are preparing to study medicine or biomedical sciences. Two specific examples outline how the draft specification was amended to meet these principles. Electricity was identified by the senior examiner for physics as a core area common across multiple GCSE syllabuses. The academic panel agreed that the topic should be retained on the BMAT specification for its relevance to biomedical topics such as nerve impulses, but indicated that the sub-topic of domestic electricity be excluded. Cosmology was also excluded from the BMAT draft specification for physics, despite frequent representation on GCSE programmes, but relevant concepts such as the Doppler effect and line absorption spectra were

retained and categorised under the topic areas of waves and wave behaviour, and the electromagnetic spectrum respectively, as medical tutors argued they were of central relevance to undergraduate study. In line with the objective of BMAT to test students' understanding of scientific principles rather than specific facts, this categorisation both ensured that candidates would not invest time revising topics that are not relevant to medical study, and encourages the abstraction and generalisation of principles beyond the initial context in which they were encountered, which is a key feature of advanced scientific reasoning.

Specific examples for core topics that would appear in the BMAT specification were also agreed. For example, homeostasis featured on most GCSE biology syllabuses but a variety of examples were used to teach the principle (e.g. regulation of blood glucose, temperature, or water content). Prior to revising the specification, only very general questions on homeostasis could be used in BMAT to avoid disadvantaging students who had encountered different examples. By including a core set of examples in the revision of the BMAT specification, a wider range of high-quality test questions could be generated.

Trial by BMAT question writers and international experts

The revised specification was shared with BMAT item writers, who assessed it again for omissions based on their experiences of writing test questions. In particular, they were asked to check that it afforded enough coverage to allow creation of high-quality questions. Feedback was very positive, and the overall opinion was that as the new specification was more explicit on the topics that could be examined, it presented new opportunities to devise challenging high-quality questions in subject areas that were previously not possible.

One aim of the revision was to support overseas students' preparation. Science and mathematics curricula internationally may place different emphasis on certain topics or techniques. Variation in question difficulty on the basis of candidate nationality presents a type of construct-irrelevant variance that test providers should seek to limit. Education professionals with a background in biomedical sciences, including teachers and university faculty staff from the Netherlands, Malaysia and Singapore reviewed the new specification to confirm that content was targeted at the level expected of students in their final year of school study in their respective countries. Creating a more detailed specification for BMAT Section 2 and defining the expected content knowledge for the test supported international students by allowing them to identify specifically where they may need to focus revision and highlighting particular areas for study.

Development of an Assumed Subject Knowledge guide

Creating a more detailed specification for BMAT also had other potential benefits for test takers. Updating and expanding the detail of the content specification provided opportunities for producing free revision materials for BMAT, because a blueprint of the content knowledge examined by BMAT Section 2 had been developed.

As the revisions to the Section 2 specification proceeded, Cambridge Assessment Admissions Testing approached Coordination Group Publications (CGP), a well-known publisher of GCSE revision guides, to collaborate on a BMAT guide, based specifically on the new specification. CGP was chosen as the preferred partner for this work as their books are visual with a minimum of text, and well known to many students and schools already, to reinforce the idea that this section of BMAT requires *revision* but not significant commitment to new learning for test takers. BMAT assessment managers reviewed CGP guides for GCSE subjects and selected pages relevant to the BMAT specification, before editing their selections to remove duplication or superfluous material. CGP commissioned their authors to produce some text and illustrations for BMAT topics that did not appear in their current books. The revision guide was compiled and made available as an online e-book, which prospective candidates can access by registering online for free.

Discussion

This case study demonstrates some of the steps that can be taken to specify a body of knowledge examined by a test, which may inform the work of others designing assessments with a knowledge component. Of course, the approach adopted for any assessment must consider the context and candidature that is being assessed. For example, a lecturer designing the assessment for a specific course they teach may find that a process similar to the one above is overly laborious. However, the principles guiding the case study should be considered in most assessment contexts. For example, consulting with colleagues responsible for more advanced courses, or even clinical placement supervisors, would help to identify crucial topics and distinguish them from those that might be considered less relevant. Simply listing the topics and sub-topics that might be included will give an idea of the scope of knowledge being assessed, which can influence decisions about the most suitable form of assessment.

Once established, a content specification can be used to consider various aspects of context validity. Although many issues that impact on BMAT's validity have been investigated using trials and research studies, some of which have been presented in previous chapters, it is also important to evaluate them when producing papers. The processes and checks involved in checking test content are presented in the following section, including the checks that compare items to the Section 2 specification.

Question paper production process

Constructing a BMAT paper is a multi-stage, iterative process which ensures that items are high quality and that each element of the test construct is represented appropriately. From the time that questions are commissioned to their appearance on the exam paper, each question goes through a rigorous process of checking and editing to ensure it is appropriately targeted to the ability of the candidates, that the topic maps to the content specification, that there are no flaws in the logic or reasoning of the question or the answer options, and that it reflects the types of general or subject-specific thinking skills stated as important by the BMAT specification.

These checks are conducted by subject matter experts (SMEs) and coordinated by assessment managers, who are typically recruited by Cambridge Assessment Admissions Testing for a combination of their subject area expertise and experience of educational contexts. A number of these assessment managers hold higher research degrees in relevant subject areas, and have experience of teaching their subject in school or university settings. On joining the organisation, they are then trained on assessment principles and interpretation of psychometric indices. Therefore, assessment managers working on admissions tests are regarded as another layer of SMEs that add to the reviews conducted by assessment experts working as consultants.

Review by SMEs has traditionally been described as a method of ensuring content validity in occupational test settings (Lawshe 1975). Content validity is concerned with the relevance of the assessment, and its items, to the targeted construct, and the degree to which it is adequately representative of the construct (Haynes, Richard and Kubany 1995); however, the term has been controversial and Fitzpatrick (1983:3) argued that 'content validity is not a useful term for test specialists to retain in their vocabulary'. Due to the focus of this volume, the technical and historical debate on content validity is not discussed fully here. Instead, we align the checks performed by SMEs on BMAT with context validity as conceptualised in the socio-cognitive framework, which includes content validity when discussing features of the task (Weir 2005).

A range of SMEs review tasks during the question paper production process. The checks focus on specific features of the task, such as the length of text input, the linguistic complexity of instructions and the time expected to be available for completing the item. Importantly, the knowledge needed to answer the task is also considered as part of the checks. For Sections 1 and 3 this is considered against a standard of everyday knowledge, whereas Section 2 items are checked against a detailed maths and science specification.

The outlines in the next sections present processes as they currently stand. Efficient and rigorous procedures have taken a great deal of time and experience to develop, so it should be noted that the sequence of reviews and checklists of issues to consider have been honed over a decade, and will continue

to be improved over time. The number of experts available for carrying out the checks considered important by Cambridge Assessment has also varied. Where necessary, item writers, vetters and academic subject experts have been recruited from educational settings and trained in the specialist skills required for reviewing test items and papers. Therefore, no claims are being made about how long the procedures described here have been in place and whether they remain a blueprint for producing future BMAT papers. There are four main phases to the preparation of a BMAT paper for Sections 1 and 3, which are item commissioning, item editing, paper construction and paper vetting. Section 2 uses a similar process but includes an additional layer of item vetting prior to the paper construction process, which is sometimes referred to as science vetting. There are some differences in the precise checks conducted for each section due to their content, and flowcharts representing each separate process are available in Figure 4.3, Figure 4.4 and Figure 4.5. However, the main focus of each stage of review is similar across the sections; these are outlined in the next parts of the chapter.

Item commissioning

A store of items, referred to as an item bank, is maintained for each subject area and thinking skill domain. This storage is used to manage the items and any associated metadata relating to them, including whether they have been used in previous tests. This allows assessment managers to monitor the number of secure items available for upcoming tests. Additional question development for BMAT begins approximately 18 months before the test. Item writers are commissioned to submit questions in their area of expertise for review and eventual inclusion in the relevant item bank. Typically, item writers are experienced teachers and question writers for school-based qualifications. Because BMAT items target constructs and reasoning that are somewhat different from those examined in GCSEs and A Levels, item writers new to the test are trained to author items targeting the skills and knowledge specified for BMAT. Training is conducted using group workshops facilitated by assessment managers.

Items submitted by less experienced item writers are scrutinised carefully by assessment managers, who provide substantive feedback. In addition, the item writers are provided with guidance on good question writing, and checklists to ensure the quality and standard of the items, which are developed from research and operational analysis. All item writers are also given feedback regularly on how certain items have performed in recent tests (e.g. the proportion of candidates getting them correct, and how well they discriminated between candidates of different levels of ability) in order to encourage reflection on the questions they develop and anticipate problems or weakness.

Item editing

Submitted items are reviewed by BMAT assessment managers to ensure the language and formatting is appropriate and to identify any obvious flaws or concerns. Typically, assessment managers for Sections 1 and 2 check the correct answers and the solution for each item submitted by the item writer. At this stage, items that are not assessing the relevant cognitive processes are identified. For example, Section 2 assessment managers identify items that can be answered correctly without knowledge of scientific processes or principles. Similarly, items that only require a test taker to recall specific scientific knowledge or facts are rejected. Assessment managers for Section 1 check the solutions submitted by item writers to ensure they match the skills defined in the test specification, and that the correct response is not ambiguous once suitable reasoning has been applied.

A senior educationalist is recruited to act as the chair for each subject (maths, physics, biology, chemistry) or thinking skills domain (understanding argument, problem solving, data analysis and inference). For BMAT Section 3, a single chair is recruited to review submitted writing tasks. Chairs are typically teachers or lecturers with extensive experience of educational contexts, and who have held or currently hold leadership roles in education. The chair reviews each item for their area, making suggestions for changes if they believe items are too easy or difficult, or that phrasing is unclear or repetitive. Items are amended and re-submitted by item writers, and then scrutinised at an editing meeting by a team of experienced item writers. Each question is critically reviewed by every group member according to a checklist of features to ensure that they are clear, solvable in the allocated time, and if destined for the Scientific Knowledge and Applications section, that the topic is relevant to the specification and the item is scientifically sound.

Due to the multiple-choice response format of Section 1 and 2 items, the cognitive processes that might lead a test taker to select each incorrect answer can be described in terms of the miscalculations or misapplication of certain processes, for further analysis. Considering this ensures that the incorrect options, referred to as distractors, are plausible answers. Furthermore, these analyses ensure that arriving at an incorrect answer can result from a failure to successfully employ the cognitive processes targeted by each section (of course they might also be selected through guessing or misreading the question). Distractors attractive to test takers for other reasons are reviewed critically to make sure that they are not contributing to construct-irrelevant variance. For example, a Section 2 question with complicated phrasing might have a distractor that would be commonly selected by those with poor language ability. This would need to be revised, because incorrect responses on Section 2 items should indicate deficiencies in science knowledge, or ability to apply this knowledge, rather than low linguistic ability. The response option

specified as correct is then subject to further scrutiny to ensure that it is not an answer that might be arrived at by accident using poor reasoning. This is done by completing the items whilst applying common flaws in reasoning, misunderstandings of scientific concepts and miscalculations.

During the editing meeting, the assessment managers also carry out checks related to English language proficiency. Although this is conducted across all sections of BMAT, it is reviewed particularly closely with Section 1 items to ensure that:

- difficulty does not come from the way the item is expressed
- difficulty does not come from the reading load
- difficulty does not come from unfamiliar cultural assumptions.

Where an editing panel or vetter considers that the level of language is too high, there are a number of remedies that can be applied. In-text glosses or paraphrases of unfamiliar terms can be provided (if such terms cannot be avoided altogether). The tone of a passage or question can be made more neutral, or the register less formal. The length of a passage can be shortened, the syntax simplified or the density of information presented can be reduced. UK-centric names, institutions or customs can be replaced with generic equivalents. All of these measures help to ensure that construct-irrelevant variance is minimised.

Item vetting (Section 2 only)

After editing meetings, Section 2 items are submitted to an extra layer of vetting at the item level that focuses on science and maths concepts, because candidates are expected to apply subject-specific knowledge when answering them. This is completed with the support of academic subject specialists, who are active research scientists in relevant fields of science and maths. These SMEs scrutinise the knowledge underpinning each Section 2 item. In particular, they check that items:

- do not rely on scientific principles taught in secondary schools that are contradicted when a more advanced model of the phenomenon is understood
- do not become ambiguous or more difficult when one considers scientific advances beyond the scope of secondary school level science
- are not answered correctly by merely recalling advanced scientific knowledge that sits outside of the BMAT Section 2 specification.

Paper construction

After editing and vetting conducted at the item level, questions are selected from available items in the bank by the chairs, taking into account how well the questions cover the test specification and the difficulty of the paper. For

Sections 1 and 2, they also estimate the difficulty of each item and then arrange the items in order of ascending estimated difficulty. The Section 3 chair selects the tasks and judges whether they are equally difficult. All chairs also select reserve items that can be substituted if later stages of review identify material to be rejected. The first draft is reviewed by an assessment manager for each subject or subskill, who also consults other assessment managers working on that section of the test to check for repetitions between topics submitted by chairs for different subjects. The assessment managers for Sections 1 and 2 also review the overall coverage and content of the section, along with the orders proposed by the chairs. The materials across the three thinking skills domains (Section 1) or four subject areas (Section 2) are collected together to form a first draft of each paper.

Paper vetting

The draft papers are then subject to paper vetting, which differs from item vetting in its focus on an entire section together. Paper vetters are typically teachers and educationalists who have worked in general educational settings. Although they come from subject specialist areas, they tend to have a broad knowledge of educational contexts that extends beyond subject-specific education. These SMEs are trained to review the entire paper in detail to check that the items conform to the specification, that there are no errors in the wording, that the keys are correct, and that the questions and any diagrams are correctly formatted. The paper is proofread for language and conformity with Cambridge Assessment's internal style guidelines. At each stage suggested changes are referred back to the chairs for checking and amendments made before the paper progresses to the next stage of checks.

A final version of the paper is scrutinised by academics from the institutions that use BMAT, who take the paper under timed conditions, and attend an extended meeting to discuss the test content and communicate any observations or issues from their experience. This engagement with the university academics is an important step in ensuring that the using institutions can give feedback on how the current paper aligns with the test construct and in maintaining the face validity of BMAT for test takers. Following this, papers are sent to externally contracted proofreaders who have not been involved in item writing or paper construction so far. Comments from this process are reviewed by Assessment Managers, who decide whether changes are made to the paper, before finalising the sections for print.

Figure 4.3 Question paper production process for Section 1

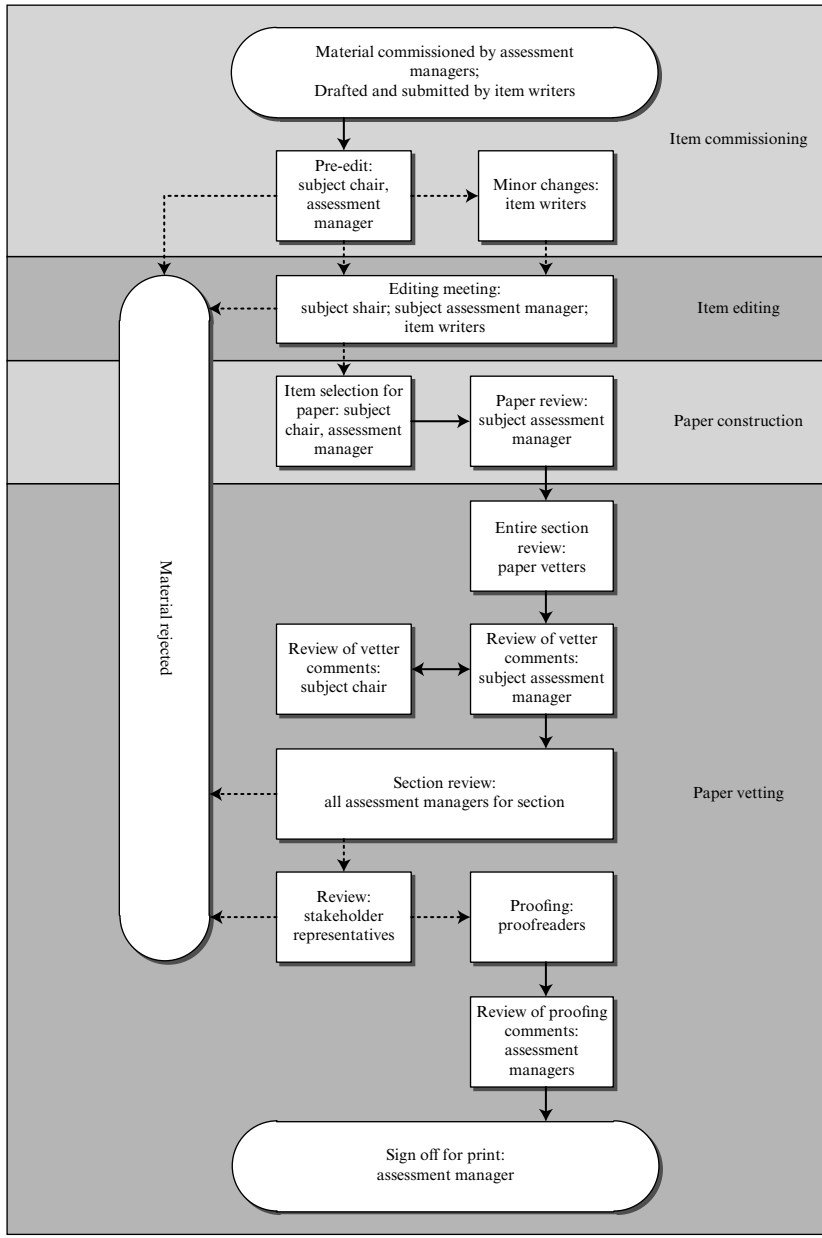


Figure 4.4 Question paper production process for Section 2

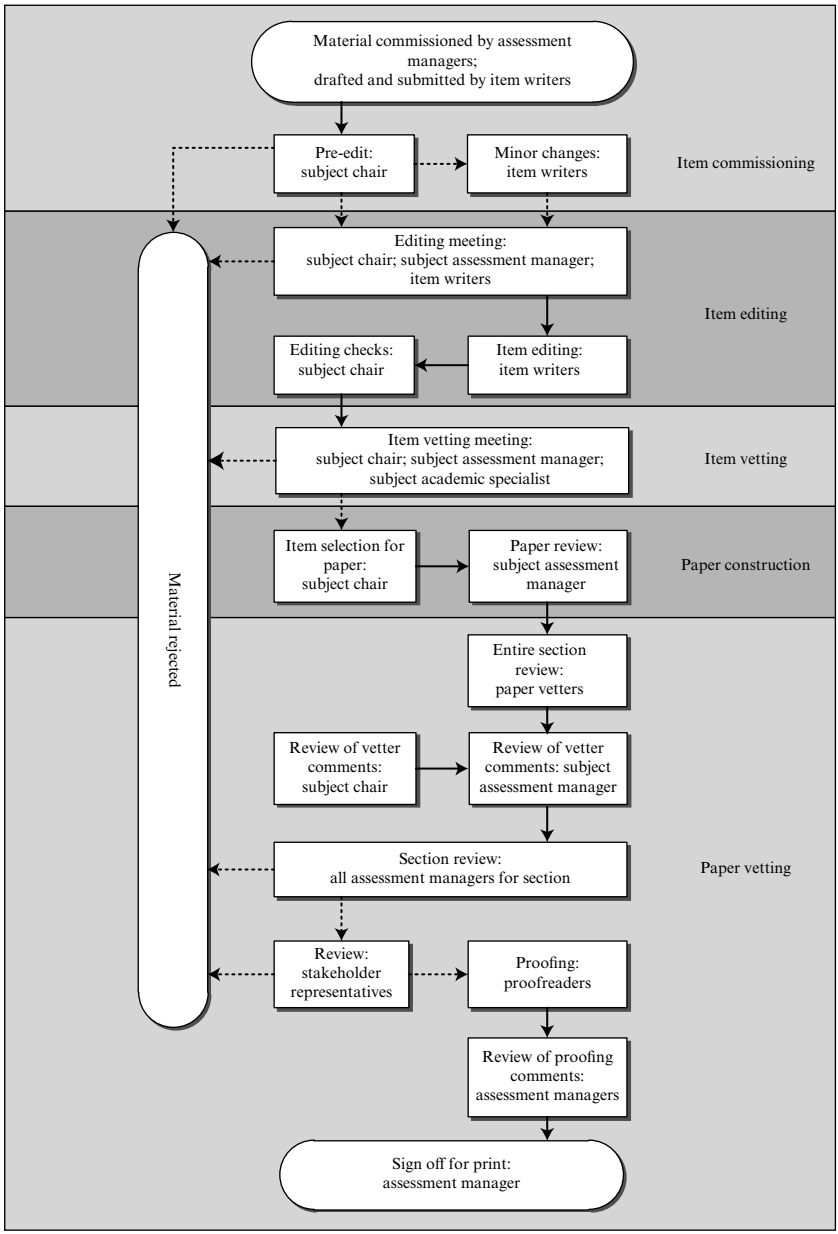
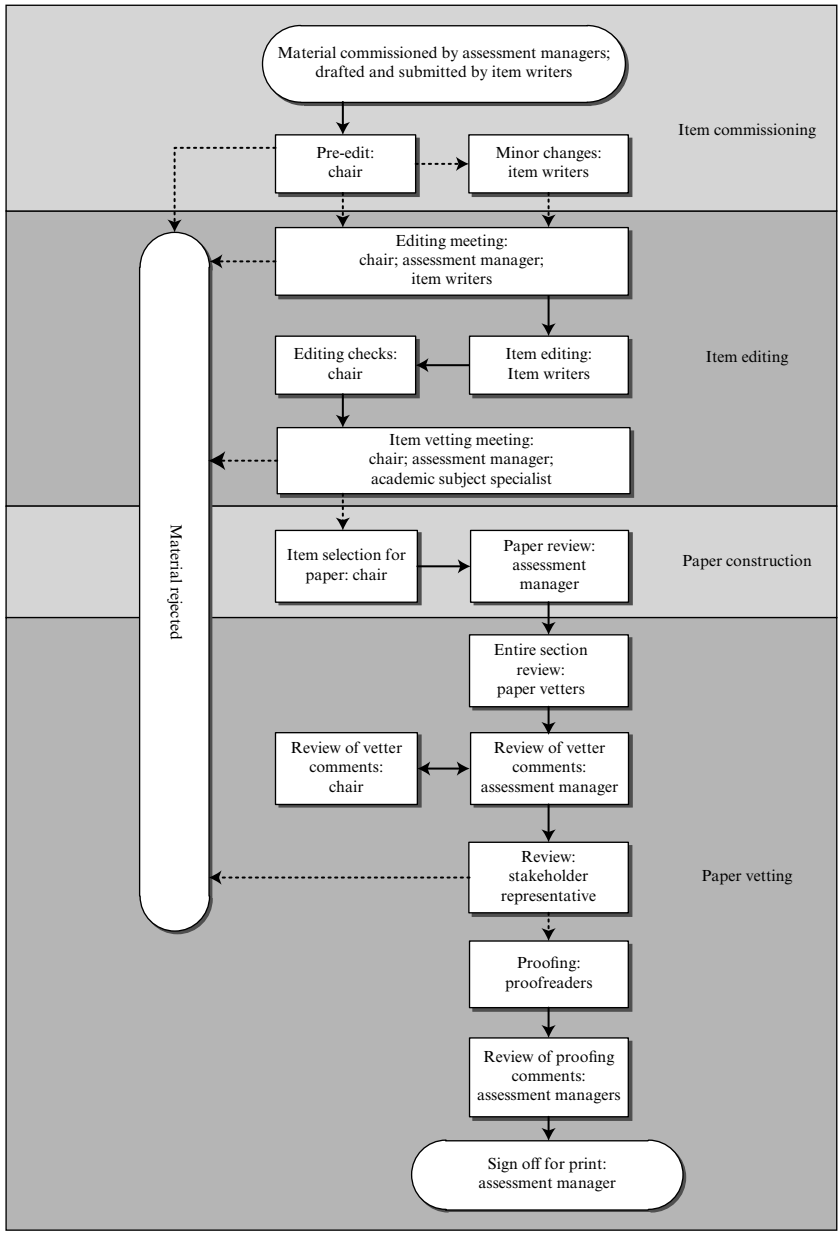


Figure 4.5 Question paper production process for Section 3



4.4 Cambridge Assessment practice: Administration features

Test delivery format

BMAT is currently paper-based (PB) for all candidates, making it accessible in test centres worldwide and unaffected by candidates' computer literacy. Furthermore, this makes it possible to administer a single test form to a large number of candidates on the same date, much like GCSEs and A Levels, with the majority of school-aged test takers completing BMAT in their own schools. At present, this would not be possible with computer-based (CB) delivery of the test due to limits in the number of computers that can be made available at the exact same time. CB would instead have to be administered across a longer time period, potentially using multiple test forms. From Cambridge English Language Assessment's experience with high-stakes English exams, we know that there are always some people looking to subvert the system. Testing windows that are open over a long period present unique challenges in this regard. Large pretested item banks and statistical methods of detecting malpractice are used very effectively to combat attempts at cheating in Cambridge English language exams; this is possible because they have larger candidatures than admissions tests and multiple sessions per year. Because these procedures are more difficult to implement for tests with smaller candidatures, it should be recognised that CB admissions testing could raise challenges to security. Specifically, a CB model would not allow every candidate to take the test on the same day, and testing over longer periods would require several test versions to be constructed as a precaution against malpractice. Furthermore, multiple language testing sessions allow malpractice panels to withhold results more readily when presented with statistical indicators of malpractice (see Chapter 5), because the test taker can retake the assessment when a result is withheld. For an admissions test that takes place once a year, such as BMAT, malpractice panels are understandably cautious when considering whether to withold results.

Whilst these security issues can be overcome with technology and multiple test forms, implementing BMAT in a CB format requires careful consideration of potential risks and threats to validity, and development of safeguards against malpractice in this context. Furthermore, medical schools using BMAT have expressed concerns that CB delivery might make it less likely that candidates could take the exam at their own schools, particularly for schools that are under-resourced in comparison to others. Despite these issues, Cambridge Assessment has been considering CB options for a number of years and consulted with BMAT stakeholders on the issue extensively. This is because CB testing has a range of advantages over PB delivery that should be recognised.

Firstly, the cost and logistical complexity of sending test papers is reduced using a CB format, and responses can be made available for operational analysis immediately. Currently with PB testing, secure delivery and return of test materials is a major undertaking that requires constant monitoring by operational teams. Secondly, contrary to the issues posed by item exposure discussed with CB admissions tests, some security issues are actually reduced when the testing organisation can control the timing of when test materials are made available more precisely, which can be achieved using technology. Currently, schools and test centres support examination boards by conforming to strict regulations governing the exam hall. Centres are instructed precisely about when packs of secure materials can be opened, to reduce the possibility that papers are exposed before a test session. Therefore, it is necessary for Cambridge Assessment to maintain a network of centre inspectors to quality assure the locations where BMAT is administered in PB form. Although CB testing would not eliminate the potential for institutional malpractice, it would likely reduce the risk from this particular threat.

The considerations outlined here are important for anyone involved in developing and delivering assessments, who might be interested in the advantages and disadvantages of PB and CB testing. It should be noted that the discussion presented here adopts the perspective of Cambridge Assessment, which is a large examining organisation that administers English exams and school-level qualifications around the world. This makes it possible to maintain inspection processes worldwide, whereas other organisations would find it difficult to accommodate this. Therefore, decisions on PB and CB assessment should include consideration of the resources available for supporting a programme of testing; this may include the financial, technological and technical assets that can be accessed readily.

Instructions to candidates

It is important that candidates can familiarise themselves with the structure of BMAT in advance so that on the test day itself their efforts are directed at performing well on the questions, rather than figuring out the format of the paper. The BMAT website provides candidates with all necessary information on the test timing, number of items, weighting of items and mark schemes, and provides sample test papers and response sheets for each test section. The mark scheme for the Writing Task is given on the BMAT website and is the same as that used by the markers themselves. This information also appears in the BMAT test specification document, which is available to candidates on the BMAT website.

In the testing situation itself, the test paper for each separate BMAT section has a front cover of instructions to candidates. The instructions state, in clear and concise terms, what candidates are required to do, the time limit,

the number of items and the marks for each item (one mark for each MCQ item). For the Section 1 and 2 test papers, candidates are instructed to work quickly (in bold), that there are no penalties for incorrect responses and (therefore) that they should attempt all questions.

For the Section 3 test paper, the instructions encourage the development of ideas, macro-planning and organising. They indicate that candidates should develop, organise and communicate their ideas, and explicitly instruct test takers to spend time thinking carefully about what they need to say (see Box 4.3).

Box 4.3 Extract from instructions on BMAT Section 3 front cover

The tasks each provide an opportunity for you to show how well you can select, develop and organise ideas and communicate them effectively in writing.

Before you begin writing, take time to think carefully about what you need to say and the ways in which the organisation and layout of your response might help convey your message. Diagrams etc. may be used if they enhance communication.

Take care to show how well you can write and be concise, clear and accurate.

These instructions are designed to support context validity by explicitly providing guidance on test-taking behaviours that give test takers the best opportunities to perform well. Providing a reminder of these in the exam hall acknowledges that candidates can forget these considerations in the pressure of a high-stakes testing environment, even if they have prepared extensively for the test.

Consistency and security of testing conditions

Cambridge Assessment provides an extensive network of test centres offering BMAT to ensure that candidates should not have to travel long distances to take the test. By guaranteeing a high degree of access to the test, institutions using BMAT can extend their own commitment to ensuring equity in access to their courses.

All candidates worldwide take BMAT under strict examination conditions. For UK-based candidates the test centre is typically in their own school or college, which is usually already a registered Cambridge Assessment examination centre running regulated examinations such as GCSEs and A Levels. This is possible because Cambridge Assessment Group includes

Oxford Cambridge and RSA Examinations (OCR), which is one of the main awarding bodies for school-level qualifications in the UK. Overseas candidates can sit BMAT in their own school or college if it has been approved by Cambridge Assessment as a suitable centre for administering high-stakes tests; applications are accepted for the approval process throughout the year. In some cases, overseas schools will already be approved by Cambridge Assessment to administer qualifications provided by Cambridge International Examinations or Cambridge English Language Assessment. Alternatively candidates may sit the test at an 'open' centre, which accepts external candidates.

Centre approval and quality assurance processes

Institutions that apply to run BMAT go through a number of checks on their suitability including storage arrangements at the premises, the availability of suitably trained invigilators and supervisors to run the tests and the centre's experience of running high-stakes international assessments. Pre-approval checks are carried out using photographic evidence, face-to-face inspections or, for some applicant centres, a 'remote' inspection using video technology may be deployed. Each application is reviewed by a number of senior managers before a decision is taken as to whether or not to authorise the institution as a centre. If there are any concerns, the centre application is declined and the candidate is directed to an alternative centre.

All Cambridge Assessment centres are provided with standardised test administration regulations. The regulations for BMAT administration are stringent and intended to ensure the highest quality in the delivery of the test, covering the secure storage, checking and return of test materials. Seating plans are also mandatory for test sessions; these record where candidates were seated in the room and in relation to each other. They are used when investigating any suspected malpractice, particularly if statistical procedures identify unusual strings of matching responses (see Chapter 5 for details). Timetable clashes, access arrangements, special considerations and the reporting of suspected malpractice are also detailed in regulations. Centres must adhere to these regulations and centre inspections are carried out to ensure their compliance. Some examples of the issues covered by regulations are described in the next parts of this chapter.

Regulations for secure storage of test materials

The centre regulations set out in detail how to transport materials securely, how to check them when received, store them safely before the test date and when exactly it is permitted to open packets of test materials. The regulations include a clear instruction to centre staff that any breach of these conditions is treated extremely seriously. An extract from the regulations is presented in Box 4.4.

Box 4.4 Extract from the Instructions for Secure Administration of Admissions Tests

Test materials should be stored in a safe; however if that is not available, a non-portable, reinforced metal cabinet with a secure lock must be used. The safe or container must be situated in a locked room, preferably windowless and on an upper floor.

The availability of secure storage is checked as part of the centre approval process and also during centre inspections. This is important because materials sometimes arrive two to three days in advance of the test date. These materials are packed so that it is not possible to remove individual papers without substantial breaches to the packaging. Exams officers at centres are instructed to store the materials securely, without opening this packaging. Centre inspectors also check that these instructions are followed and a failure to do so would be considered a serious breach of regulations.

Checking the identity of candidates

Confirming that the person who takes the test is the same one who is registered for it and who receives the result is central to safeguarding the validity of test results. Stringent identity checks are also detailed as part of Cambridge Assessment's regulation documents, which specify that candidates need to produce 'an original photographic ID, for example, a passport, national identity card, photographic driving licence etc.'. These precautions are used to avoid the possibility of imposters taking a test in place of the genuine candidate.

Monitoring centres' compliance with regulations

Test centres are regularly monitored on the quality and compliance of test delivery. These inspections are carried out by trained inspectors who visit a test venue on the test day and observe the secure storage arrangements, checking of IDs and invigilation of the test, with the aim of ensuring compliance with the regulations in the *Instructions for Secure Administration of Admissions Tests* (Cambridge English 2014). Reports detailing the centres' performance, including recommendations for improvements, are produced after each inspection. An example report extract from a centre inspection is given in Box 4.5. The inspector visited the centre and observed the secure storage and invigilation arrangements, rating them as 'Fully compliant' (i.e. no faults).

Inspections do occasionally identify centres that do not meet the standards set by Cambridge Assessment Admissions Testing. Another example report extract from a centre inspection is given in Box 4.6, but this time, the centre was rated 'Unsatisfactory'. As shown in the extract, reports provide

Box 4.5 Report from an inspection of a UK centre – November 2013

Security arrangements at the school are excellent. The [storage] room has no windows, a solid and lockable door and a security 'grill' on the outside of the door. The room is also alarmed and exam/test materials are stored in locked metal containers that are bolted to the wall. There is also an area where papers can be sorted and processed on arrival and packaged up prior to being returned to exam boards.

The Exams Officer and her team are meticulous in the planning for each exam/test.

I was with the Exams Officer when she removed the papers from the security room and transported them to the invigilator. At no time during the whole session were the papers left unattended and they were opened in front of the candidates. When materials arrive at the centre they are checked immediately and locked away in the security room.

After the exams/tests have been completed they are stored according to the ATS instruction booklet.

The room had excellent light and was warm, well ventilated and quiet, very conducive to doing an exam.

feedback for the exam centre to act upon. Whilst inspections that identify unsatisfactory centres happen rarely, instances are taken seriously. Centres receiving this rating are referred for inspection again at the earliest opportunity, and failure to address concerns can result in withdrawal of a centre's eligibility to administer exams.

Cambridge Assessment is continually looking for ways to enhance the security and quality of examination delivery. Regulations are used to ensure BMAT is delivered as uniformly as possible around the world, and centre inspections are used to maintain these standards. While costly to maintain, these measures are essential for ensuring that candidates and other test users can have complete confidence in the consistency and fairness of the assessment.

These measures are proportionate to the scale and stakes involved in taking BMAT. For individuals designing bespoke assessments or class tests, the level of scrutiny outlined here might be considered excessive. However, even small-scale examinations can have very high stakes; in these contexts those designing assessments are advised to consider the security of their administration conditions carefully. In addition, the environment for administering any assessment should be evaluated, because this is easily overlooked even though it can impact on candidates' performances. For example, the

Box 4.6 Extracts from an inspection report – November 2016

After the test the exam scripts from the test room were transported to another exam room unsealed. The inspector also noticed that the materials were left in an office briefly before being transported for packing in another room. Please ensure that test materials are always sealed before leaving the test room; this is to ensure the materials cannot be tampered with or become lost or damaged in transit.

In an exam room, the invigilator was using a mobile phone to send messages, a tablet and a laptop during the test. The invigilator also briefly left the test room unsupervised whilst going out into the corridor to monitor noise levels. Invigilators must be attentive throughout the test and must not do any other activity in the test room, for example, reading a book or working on a laptop. Candidates should also never be left unsupervised, even for a brief period of time. Where there is only one invigilator present they must be able to get help easily, without leaving the test room.

temperature in a room and its dimensions can influence the comfort of test takers, particularly if the assessment is a long one.

4.5 Chapter summary

In this chapter we have outlined the factors that constitute the context validity of a test, focusing on the tasks and administration conditions that are considered when providing an admissions test such as BMAT. Many decisions related to the initial design of BMAT were informed by research evidence, such as the inclusion of both MCQs and constructed-response tasks in the test. Studies have also been conducted by Cambridge Assessment researchers to investigate specific aspects of BMAT, such as timing allocated to sections. Importantly, this research informed improvements to BMAT's specification and also to the processes that consider context validity in item production.

The processes used to govern BMAT item writing and test administration have also been described in detail as part of the present chapter. Our practice in this area strives to safeguard cognitive validity and make the experience of sitting BMAT as fair as possible to candidates, wherever in the world they may be taking it. Worldwide access to BMAT means that there are no exemptions to taking the test, and this is an important consideration for selecting institutions that wish to have a common measure on which to compare *all* their applicants. It is important for Cambridge Assessment to demonstrate the rigour of our processes to universities using BMAT and also to test takers, who should be confident that the testing conditions they face

are controlled as far as practically possible. However, the detail included in this chapter also serves a second purpose. By outlining our approach and the issues we consider as a large examinations organisation, we hope to share experiences that are useful for those designing their own assessments. In many cases the resources available to smaller organisations will prevent the level of scrutiny outlined in the present chapter, but the examples should serve to demonstrate how context validity principles and questions posed by Weir (2005) can be applied in practice.

The present chapter has highlighted the emphasis on item and paper production that underlies BMAT, which can be considered part of *a priori* validation in Weir's (2005) framework, along with cognitive validity. Both context and cognitive validity are used to ensure that test design and production supports the ultimate goal of providing BMAT scores that are meaningful for the medical, veterinary and dentistry schools using the test to select applicants. The next chapter on scoring validity focuses more specifically on the meaningfulness of scores, and represents the first stage of *a posteriori* validation (Weir 2005).

Chapter 4 main points

- Context validity focuses on features of the tasks and features of the administration.
- For BMAT, Cambridge Assessment considers a wide range of issues that relate to task design, task construction and test administration.
- Different response formats have varying strengths and weaknesses; this can be addressed in assessments by including both MCQs and tasks requiring constructed responses.
- The exact processes used can differ depending on format, focus and scale of the test, but the underlying issues to consider can be applied to many assessment contexts.

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.