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

3 What skills are we assessing? Cognitive validity in BMAT

Kevin Y F Cheung

*Research and Thought Leadership Group,
Cambridge Assessment Admissions Testing*

Sarah McElwee

*Research and Thought Leadership Group,
Cambridge Assessment Admissions Testing*

3.1 Introduction

This chapter focuses on theory-based validity (Weir 2005), which has more recently been referred to as cognitive validity (Field 2011), for the three sections of BMAT. Cognitive validity refers to the cognitive processes engaged by test takers when attempting test tasks, and the degree to which they resemble the processes engaged in non-test settings. Therefore, establishing cognitive validity requires a test developer to consider the rationale for measuring particular constructs and relevant cognitive theories of how the processes function.

Cognitive validity provides the rationale for selecting which constructs to measure and the theoretical underpinnings of these constructs; context validity is concerned with the tasks and administration conditions used to elicit these constructs; while scoring validity provides the evidence for how effectively and accurately the constructs were measured. The three together have a symbiotic relationship (O'Sullivan and Weir 2011) and can be thought of as overall construct validity (Weir 2005), which is whether the test assesses what it purports to measure, and also whether it is fit for its intended use. Terms such as aptitude, ability and skill are used to describe the constructs assessed by admissions tests, and many researchers commonly refer to admissions tests as aptitude tests (e.g. McManus, Powis, Wakeford, Ferguson, James and Richards 2005); however, this characterisation of admissions tests has been criticised and contested by others (Bell, Judge, Parks, Cross, Laycock, Yates and May 2005, Jencks and Crouse 1982), because aptitude is often interpreted as referring to innate ability. Furthermore, there have been historical changes to test specifications and theories underlying the use of particular terms, which impact on how they are used and understood by researchers. For example, the SAT, a college admissions test widely used in the US, was originally called the Scholastic Aptitude Test, but this was then changed to the Scholastic Assessment Test (Newton and Shaw 2014). These days SAT is no longer an acronym and avoids the complicated connotations associated with various terms.

To support discussions in this chapter and throughout the rest of the volume, definitions of key terms are presented in Table 3.1. These are informed by discussions in the psychometric and educational assessment literature (in particular Kaplan and Saccuzzo 2012, Newton and Shaw 2014, Stemler 2012); the definitions presented here are used by Cambridge Assessment researchers working on admissions tests, and it is acknowledged that they may not be universally accepted.

Table 3.1 Definitions of key terms

Term	Definition
Ability	The current level of performance, as contributed to by a combination of innate characteristics, academic study and individual preparation. Ability can be assessed in domain-general and domain-specific contexts.
Achievement	Competence in an area, normally subject specific, demonstrated through prior attainment of a qualification, such as A Level grades.
Aptitude	The potential for developing a skill, based on innate characteristics.
Domain-general measure	An assessment of general thinking, reasoning or problem solving skills that could be applied in a number of different contexts and subject areas.
Domain-specific measure	An assessment linked to learning information from a specific content area. The area can be explicitly defined by a curriculum or indicated by specifying a topic area.
Intelligence	Intelligence is a contested term, without a common definition. Therefore, Cambridge Assessment avoids use of the term in test specifications, which should be unambiguous. Many applications of intelligence refer to innate characteristics, although this is controversial. Cambridge Assessment researchers do sometimes refer to intelligence, and the theory of intelligence being used is explicitly referred to in such cases.
IQ (intelligence quotient)	An expression of an individual's intelligence, as measured at a particular time by a specific instrument.
Knowledge	Information about processes or topics that can be codified and learned. In particular, knowledge refers to sets of facts that can be memorised and recalled.
Potential	Having or showing the capacity to develop into something in the future.
Skill	An ability that can be progressively developed, often through learning and practice.

An important issue for discussing the cognitive validity of an admissions test is the distinction between aptitude, ability and achievement. Cambridge

Assessment researchers have previously used aptitude as a synonym for potential (Emery and Bell 2011) and BMAT Section 1 is titled Aptitude and Skills to reflect this original intended meaning. However, referring to aptitude in the test specification is currently being reviewed by Cambridge Assessment Admissions Testing in order to align the term's usage with conventions established in educational psychology. The term aptitude has historical connotations linked to assessing innate abilities, and to assessing special aptitudes as part of vocational guidance (Newton and Shaw 2014). In recognition that aptitude is not used merely as a synonym for potential by some contemporary researchers (see Box 3.1 for an example), the term is not used to describe BMAT's test construct in this volume, except where discussing Section 1's full title.

Box 3.1 Stemler (2012:11) on the distinction between aptitude, ability and achievement

For one individual, it may take a decade to learn to perform a particular piece of music, and for another individual it may take only a few days. The latter would be said to have higher *aptitude* than the former; however, both individuals share the same degree of *ability* in that they demonstrate with equal competence the mastery of a specific skill set . . . . A person with high music ability who has never participated in a concert or been evaluated by a teacher as achieving a particular "grade" level on an instrument lacks demonstrated *achievement*, even as she may possess high ability and/or aptitude. (emphasis added)

For BMAT, which focuses on combining knowledge with successful application of skills rather than mere recall, the cognitive processes targeted by the assessment form the core of the test construct. In contrast, for a test of pure knowledge, context validity would form a larger proportion of the test construct, because it includes consideration of the content knowledge required to complete tasks; whereas cognitive validity would primarily consider cognitive models of memory retrieval.

Deciding which cognitive processes to assess are key decisions made during the planning and design phases of the test development cycle (see Chapter 1 for an overview of the phases). Ongoing review of cognitive validity is also of interest for test providers, to ensure that the test assesses the skills or abilities that it is intended to measure. Questions that can be answered through trivial means, such as eliminating entirely implausible options or through unintended clues or using extraneous information, compromise the validity of an assessment. Achieving a high score on a test of thinking skill, scientific reasoning, or written communication should require candidates to engage in cognitive processes similar to those they would use in relevant real-world contexts. This illustrates that cognitive validity does not exist in isolation from the test's use; instead, consideration of how cognitive skills will be used after the test shapes the design of the assessment.

Furthermore, contextual features of a task, such as the time allocated, response format, and dependency on prior knowledge are determined by the cognitive processes that the test items and tasks are intended to assess. By outlining these issues, the present chapter demonstrates how consideration of cognitive validity and context validity is intertwined for BMAT, and argues that all test providers have a responsibility to consider both cognitive and context validity.

Researchers face specific challenges when exploring cognitive validity, partly due to the nature of standardised assessment. Generally, a test score is derived from assessing a final submission, whether a set of selected responses or a candidate-constructed response (e.g. an essay). Therefore, research based on test scores, and in fact interpretation of scores for selection purposes, rely on inferring that correct or high-scoring responses result from the cognitive processes being targeted. Similarly, a low score indicates that the test taker did not perform the targeted cognitive process to a high standard, but live test sessions rarely offer direct evidence of this inference. This means that *a posteriori* data from live test sessions can only provide limited evidence of cognitive validity; therefore, *a priori* theories about the test construct form the basis of cognitive validity (Weir and O'Sullivan 2011).

The present chapter outlines how cognitive features were considered in the development of BMAT from predecessor assessments and how the cognitive processes of test takers have been investigated with research. In order to contextualise the application of Weir's (2005) framework, the following part of the chapter focuses on cognitive validity specifically in relation to BMAT.

3.2 Cognitive validity and its importance to BMAT

Correctly answering a test task should require candidates to replicate cognitive processes which might be required in relevant non-test contexts, which in the case of BMAT are the study of medicine, dentistry and related subjects. The focus on biomedical study, rather than clinical practice, is a conscious decision that has important implications when considering the cognitive validity of BMAT. This approach recognises that universities may evaluate an applicant's suitability for the actual practice of medicine and dentistry, by employing other selection methods. Furthermore, some of the non-academic skills used in clinical settings are specific to the context and targeted for development as part of clinical training. Therefore, investigation of BMAT's cognitive validity is concerned with the relevance of the assessed skills for biomedical study, and also how these skills will be employed in the testing environment compared to the real world. In the context of high-stakes tests, a compromise is inevitably needed between considering task authenticity (i.e. the degree to which they resemble tasks faced in non-test environments) on the one hand, and safeguarding against other threats to validity on the other. Increasing the authenticity of a test at the cost of scoring validity, fairness, or security would typically not be acceptable from an overall validity argument perspective.

One approach for achieving task authenticity is to design tasks that simulate how the assessed skills will be used. For example, a speaking exam might role play situations that a test taker would expect to encounter when using the language they have learned, such as shopping or asking for directions (Galaczi and ffrench 2011). However, this is more difficult for an admissions test such as BMAT, because medical study is something that a BMAT candidate is unlikely to have experienced before sitting BMAT. Therefore, performance on an authentic and relevant task can depend on knowledge that the test taker would not reasonably be expected to have at the point of applying. An authentic task with good cognitive validity for medical study might require the test taker to complete an exam testing advanced knowledge of physiology. Alternatively, a task assessing skills used in clinical practice would ask the applicant to conduct a medical procedure or a differential diagnosis for a simulated patient. Although these would elicit relevant cognitive processes in some test takers, it is unreasonable to assume that all applicants to medical school have the necessary knowledge to complete these tasks successfully, particularly as these skills are intended to be taught as part of the course for which they are applying. In addition, more specialised tasks can be particularly difficult to prepare for, and there can be differences in the availability of preparation materials or access to work experience opportunities in the medical profession. These tasks could potentially compromise other aspects of validity, illustrating how cognitive validity is best considered alongside test taker characteristics and consequential validity, which includes issues of fairness and bias.

Given these concerns, it is clear that the cognitive processes and skills to be assessed should be the subject of careful consideration in a medical selection context. According to Weir (2005), cognitive validity poses a main research question: What are the skills/cognitive processes elicited by the test tasks? Before investigating this question however, there is another one posed by cognitive validity that is particularly relevant to BMAT: What are the skills/cognitive processes that the test **should** aim to elicit?

For BMAT, cognitive validity can be conceptualised as the extent to which test items elicit the types of mental skills required of a biomedical student during their course of study. Clearly, there are determinants of successful biomedical study that are beyond the scope of what can be validly assessed with an admissions test. For example, personal qualities and personality traits, sometimes referred to as non-cognitive skills (Patterson, Knight, Dowell, Nicholson, Cousans and Cleland 2016), are increasingly recognised as important in the context of medical study and practice (Katz and Vinker 2014, Koenig, Parrish, Terregino, Williams, Dunleavy and Volsch 2013, Powis 2015); these might not be suitable for testing in the exam hall. Therefore, admissions tests such as BMAT should be seen as one tool in the selection process that is recommended to be used alongside other methods of assessment and evaluation (Cleland et al 2012). Standardised testing can only claim to assess a subset of the criteria that might facilitate successful medical study, and cognitive validity should inform decisions regarding all assessments that form part of this process. Part of the test developer's responsibilities for cognitive validity lie in considering the constructs and processes most relevant to the non-test setting that is being selected for. This is not a simple task in the context of biomedical study. Even if a test developer only concentrates on cognitive abilities, the constructs that might be described as relevant to biomedical study are infinite, because constructs can be broadly or narrowly defined.

A case could be made for testing numeracy, working memory capacity, knowledge of statistics, verbal reasoning, or even spelling ability. However, the strength of the theory and evidence base for the relevance of these constructs must be evaluated as part of test design. Due to the range of plausible constructs to assess, a key decision for the test developer is the selection of the ones which are most suitable. Given unlimited testing time and applicants who are immune to fatigue, a multitude of scores could be produced for various skills, but this is not a practical starting point for a test provider. Almost all of the attributes that might be tested include a cognitive component, even those referred to as non-cognitive; for example, responding to a self-report personality assessment typically requires candidates to engage reflective processes. It is these cognitive processes engaged in responding that are considered when evaluating cognitive validity.

To address both the question of what skills should be assessed and which processes are actually elicited by BMAT, the main content of this chapter is split into two parts. Firstly, in part 3.3, the original impetus for the test and its roots in predecessor assessments are described to contextualise the constructs measured by each section of BMAT, with a focus on their cognitive components. The construct of a test is what the test purports to measure, including the cognitive theory (or theories) regarding the skills being targeted by the test. Relevant literature on the cognitive processes involved in critical thinking, problem solving, scientific reasoning and written communication are briefly reviewed, alongside example BMAT tasks. The rationales for selecting these skills are discussed, presenting an argument that these are the skills that should be assessed for biomedical study. Discussion of how cognitive validity intertwines with other aspects of validity is also included throughout this part of the chapter, particularly with regard to Section 2 and context validity.

The next portion of the chapter, part 3.4, provides examples of Cambridge

Assessment research studies that evaluate how well BMAT tasks assess the identified skills. These illustrate approaches to investigating cognitive validity that can be applied to other similar tests, and how the findings can inform the overall evaluation of a test.

3.3 Selecting and defining the cognitive processes included in BMAT

Deciding on the cognitive processes to assess is an important component in the design and planning phases of a test development and validation cycle (see Chapter 1 for an overview of the cycle). It is vital that the cognitive processes targeted by a test are well defined, so that papers are constructed to capture these qualities suitably.

BMAT has its origins in a programme of collaborative research and development involving the University of Cambridge, the University of Oxford, University College London (UCL), Imperial College London and the Royal Veterinary College (RVC). The test was designed to supplement existing sources of information (such as examination results, personal statements and performance at interview) to aid the process of selection for competitive biomedical degree courses, and the kinds of reasoning assessed by BMAT reflect this.

The institutions involved in developing BMAT had a number of requirements in common:

- to differentiate between applicants with the highest prior attainment in their school examinations
- to ensure that applicants' scientific understanding is adequate for the study of biomedical sciences, and that they can cope with the demands of a rigorous science-based course
- to provide a common measure for comparing applicants from a variety of educational backgrounds and with a variety of qualifications, including overseas applicants, mature applicants, and applicants from different school types, many of whom only had predicted grades at the point of application
- to allow admissions staff to focus resources towards applicants with a realistic chance of receiving an offer.

Precursors to BMAT

Cambridge Assessment's involvement in developing tests of academic aptitude stretches back to the 1980s with the development of a Law Studies Test in collaboration with the Law Schools Admissions Services in the United States (Black 2012). This, together with further work done by Alec Fisher on a proposed test of academic aptitude for higher education (Fisher 1990a, 1990b) led to a wider project, MENO, that set out to identify, define and assess those thinking skills that are important for success in higher education (Chapman 2005). At its conclusion, the MENO project also developed standardised assessments of these skills for use in higher education selection; therefore the tests from MENO are ancestors of various admissions tests that have a domain-general thinking skills component, including the Thinking Skills Assessment (TSA) and BMAT.

The development of BMAT was influenced by the findings of the MENO project and the requirements of the various universities outlined previously. BMAT's development was also informed by two pre-existing tests that had been trialled and shown to make a positive contribution to student selection: the Cambridge Medical and Veterinary Admissions Test (MVAT) and the Oxford Medical Admissions Test (OMAT) (Emery and Bell 2009, James and Hawkins 2004).

BMAT's test construct

The rationale and construct for each of the BMAT sections, and the influence of the original MVAT and OMAT tests in the development of BMAT is outlined in the following overview. Brief descriptions of relevant theories, taxonomies and models have also been included to contextualise and promote a theory-based approach to validity (Weir 2005).

Section 1 – Aptitude and Skills

The Aptitude and Skills section is designed to assess candidates' thinking skills, which can be thought of as specific cognitive abilities. BMAT Section 1 includes three types of item: problem solving, understanding argument (sometimes known as critical thinking) and data analysis and inference. Early conceptions of the questions that now form the basis of BMAT Section 1 BMAT come from Fisher's (1992) work on the higher education aptitude tests and the MENO thinking skills project. This early work identified a number of areas that became components of overall thinking skills. For example, Fisher proposed the construct of logical reasoning as a precursor to understanding argument, which should test:

> [T]he kinds of reasoning skills which are used in everyday arguments (i.e. arguments which . . . have been actually used by authors with a view to persuading their readers). [Questions are] expressed in natural language and do not use symbolic languages (or symbolic logic). Stimulus passages contain some reasoning or they contain sufficient subject information to serve as a basis for argument. The subject matter of logical reasoning items ranges very widely and may include anything from cigarette smoking . . . or natural science to law (Fisher 1992:4).

Fisher also lists a number of core skills identified by various education and curriculum bodies as central to successful critical thinking, and representative items from the understanding argument part of BMAT Section 1 comprise:

- summarising the main conclusion of an argument
- drawing a conclusion when premises are given
- identifying assumptions
- assessing the impact of additional information
- detecting reasoning errors
- applying principles.

The MVAT pilot at University of Cambridge in 1999 drew heavily on the research and development work of the MENO project. This, in turn, influenced the content of BMAT Section 1, which includes item types developed and refined during the MENO project. The design of BMAT Section 1 was also influenced by work conducted at the University of Oxford's medical school.

James and Hawkins (2004) describe a review of selection processes at Oxford to explore the range of practices and use of test scores for the university's internal selection test for medicine. Despite a variety of selection practice across the colleges of the university, it was possible to distil key abilities that were highly rated by tutors as follows:

- understanding of written texts, particularly extracting meaning from complex work
- understanding numerical data and the representation in graphical form, including extracting meaning from datasets
- communication through the use of clear written English to express abstractions and arguments
- ability to use diagrams, graphs and text to express results and arguments
- thinking at an abstract and conceptual level, including logical and numerically based reasoning.

The resulting BMAT Section 1 specification was derived from early pilots and includes three skills considered important for successful study in higher education. As these skills are beneficial across many subject areas, some of the item types are similar to those included in more general assessments. The definitions of the skills included in the BMAT test specification are presented in Box 3.2.

The description of each item type in the test specification (problem solving, understanding argument or data analysis and inference) outlines the subskills that a test taker must employ to answer test items correctly, rather than the knowledge that they need to demonstrate; in other words, these are the cognitive processes assessed by Section 1. For each item type, the definition is restricted to aspects of the skill that are relevant for higher education study and also suitable for standardised assessment.

Box 3.2 The BMAT Section 1 test specification

Problem solving
Demands insight to determine how to encode and process numerical information so as to solve problems using simple numerical and algebraic operations. Problem solving will require the capacity to:

- select relevant information
- recognise analogous cases
- determine and apply appropriate procedures.

Understanding argument
Presents a series of logical arguments and requires respondents to:

- identify reasons, assumptions and conclusions
- detect flaws
- draw conclusions.

Data analysis and inference
Demands the use of information skills (vocabulary, comprehension, basic descriptive statistics and graphical tools), data interpretation, analysis, and scientific inference and deduction to reach appropriate conclusions from information provided in different forms, namely:

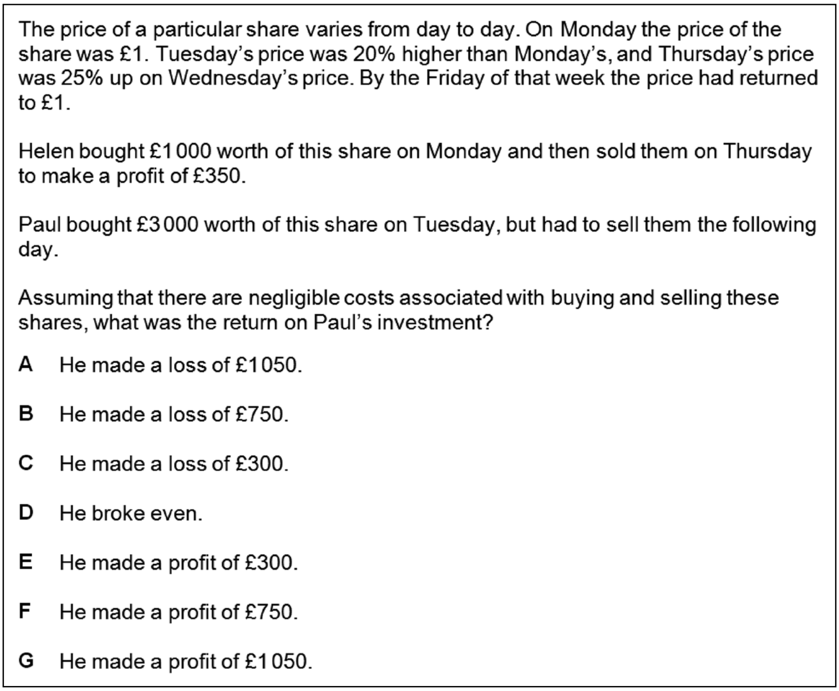
- verbal
- statistical
- graphical.

(Admissions Testing Service 2016b)

Problem solving is included in other Cambridge Assessment qualifications and examinations, such as TSA. The Cambridge International A Level in Thinking Skills provided by Cambridge International Examinations also contains a problem solving component and defines it in the syllabus as: ‘a candidate’s ability to analyse numerical and graphical information, which is based in real life situations, and apply the right numerical techniques to find new information or derive solutions’ (Cambridge International Examinations 2016:9).

Early versions of the test task specifications that eventually became problem solving were trialled as part of MENO and described as mathematical reasoning, but this was changed to formal reasoning, referring specifically to reasoning in a mathematical context. Following a review of the trials and the cognitive processes it would be desirable to elicit, the label for the category was changed to problem solving, in order to reflect a renewed focus on dealing with novel problems presented in numerical, graphical and spatial contexts. An example problem solving item is presented in Figure 3.1.

Figure 3.1 Example problem solving (finding procedures) item from BMAT 2015



The price of a particular share varies from day to day. On Monday the price of the share was £1. Tuesday's price was 20% higher than Monday's, and Thursday's price was 25% up on Wednesday's price. By the Friday of that week the price had returned to £1.

Helen bought £1 000 worth of this share on Monday and then sold them on Thursday to make a profit of £350.

Paul bought £3 000 worth of this share on Tuesday, but had to sell them the following day.

Assuming that there are negligible costs associated with buying and selling these shares, what was the return on Paul's investment?

A He made a loss of £1 050.

B He made a loss of £750.

C He made a loss of £300.

D He broke even.

E He made a profit of £300.

F He made a profit of £750.

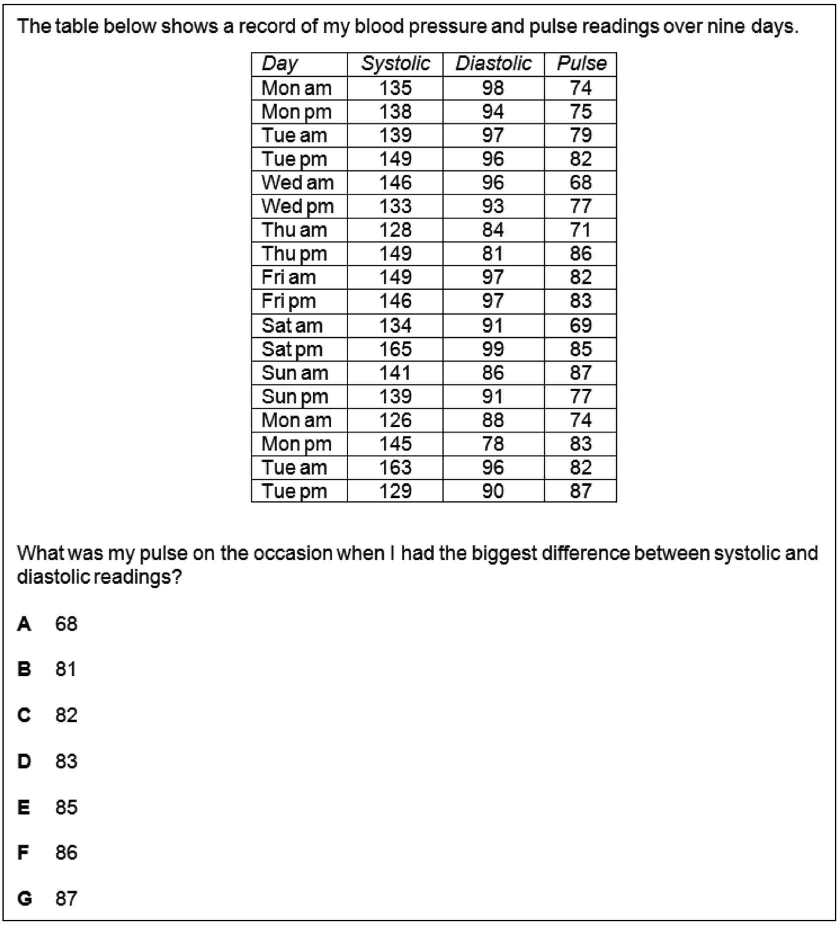
G He made a profit of £1 050.

To solve this problem, the candidate must evaluate the figures provided and find the procedure that will give the correct answer. It is not immediately clear what calculations need to be conducted and the test taker must think at an abstract level to discover what can be done with the information that is available. Once the correct procedure has been identified, the calculations are not difficult to carry out, because the item is not designed to assess mental arithmetic; instead, the question targets the ability to identify innovative solutions to the problem. This is supported by the incorrect response options available to the candidate, which are suitable as they are arrived at by following an incorrect procedure, rather than from making mistakes in the calculations.

Figure 3.2 shows another problem solving item, which asks the test taker to evaluate a table with over 70 cells and over 50 values. This problem requires the candidate to select the information relevant for answering the question. Identifying the correct answer as F does not require the test taker to carry out complex calculations.

Another category of items in BMAT Section 1 is the understanding argument items. Understanding argument focuses on logical reasoning and is

Figure 3.2 Example problem solving (relevant selection) item from BMAT 2016



The table below shows a record of my blood pressure and pulse readings over nine days.

Day	*Systolic*	*Diastolic*	*Pulse*
Mon am	135	98	74
Mon pm	138	94	75
Tue am	139	97	79
Tue pm	149	96	82
Wed am	146	96	68
Wed pm	133	93	77
Thu am	128	84	71
Thu pm	149	81	86
Fri am	149	97	82
Fri pm	146	97	83
Sat am	134	91	69
Sat pm	165	99	85
Sun am	141	86	87
Sun pm	139	91	77
Mon am	126	88	74
Mon pm	145	78	83
Tue am	163	96	82
Tue pm	129	90	87

What was my pulse on the occasion when I had the biggest difference between systolic and diastolic readings?

A 68

B 81

C 82

D 83

E 85

F 86

G 87

sometimes referred to as critical thinking, as in the *BMAT Section 1 Question Guide* (Admissions Testing Service 2016a). They are informed by Cambridge Assessment's extensive work on assessing and operationally defining critical thinking as part of MENO, which is summarised by Black (2012). As a result, the understanding argument items assess elements of critical thinking identified in Black's (2008) taxonomy, particularly the analysis and evaluation skills (see Table 3.2 for subskills).

Compared with the breadth of critical thinking subskills assessed in TSA, BMAT Section 1 includes a more limited set of items that focus on analysing and working with logical arguments expressed in everyday language, hence

Table 3.2 Taxonomy of critical thinking skills included in Cambridge Assessment's examinations (Black 2008)

Skill/process	Subskill/Sub-process
1 Analysis	A Recognising and using the basic terminology of reasoning
	B Recognising arguments and explanations
	C Recognising different types of reasoning
	D Dissecting an argument
	E Categorising the component parts of an argument and identifying its structure
	F Identifying unstated assumptions
	G Clarifying meaning
2 Evaluation	A Judging relevance
	B Judging sufficiency
	C Judging significance
	D Assessing credibility
	E Assessing plausibility
	F Assessing analogies
	G Detecting errors in reasoning
	H Assessing the soundness of reasoning within an argument
	I Considering the impact of further evidence upon an argument
3 Inference	A Considering the implications of claims, points of view, principles, hypotheses and suppositions
	B Drawing appropriate conclusions
4 Synthesis/ Construction	A Selecting material relevant to an argument
	B Constructing a coherent and relevant argument or counter-argument
	C Taking arguments further
	D Forming well-reasoned judgements
	E Responding to dilemmas
	F Making and justifying rational decisions
5 Self-reflection and self-correction	A Questioning one's own preconceptions
	B Careful and persistent evaluation of one's own reasoning

the label 'understanding argument'. Despite this narrower focus, the items include the core components of critical thinking and explicitly target processes involved in rational thought. This aligns the items with the Cambridge Assessment definition of critical thinking presented by Black's (2008) work using expert consensus (Box 3.3).

An example understanding argument item from BMAT 2016 is presented in Figure 3.3. This item asks the test taker to read a passage and identify the assumption underlying the argument that is presented in the text.

To identify that D is the assumption, the candidate must understand the argument made in the passage, and then evaluate the response options in relation to this understanding. Another understanding argument item is

Box 3.3 The Cambridge Assessment definition of critical thinking

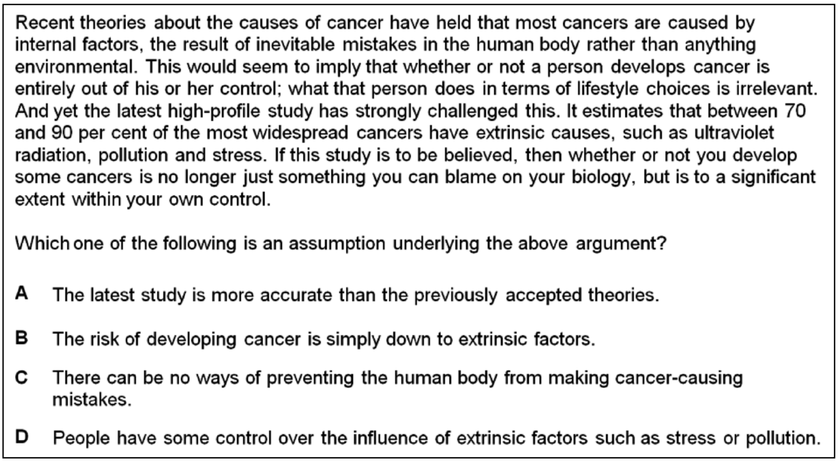
Critical thinking is the analytical thinking which underlies all rational discourse and enquiry. It is characterised by a meticulous and rigorous approach. As an academic discipline, it is unique in that it explicitly focuses on the processes involved in being rational. These processes include:

- analysing arguments
- judging the relevance and significance of information
- evaluating claims, inferences, arguments and explanations
- constructing clear and coherent arguments
- forming well-reasoned judgements and decisions.

Being rational also requires an open-minded yet critical approach to one's own thinking as well as that of others.

(Black 2008)

Figure 3.3 Example understanding argument (identifying assumptions) item from BMAT 2016



Recent theories about the causes of cancer have held that most cancers are caused by internal factors, the result of inevitable mistakes in the human body rather than anything environmental. This would seem to imply that whether or not a person develops cancer is entirely out of his or her control; what that person does in terms of lifestyle choices is irrelevant. And yet the latest high-profile study has strongly challenged this. It estimates that between 70 and 90 per cent of the most widespread cancers have extrinsic causes, such as ultraviolet radiation, pollution and stress. If this study is to be believed, then whether or not you develop some cancers is no longer just something you can blame on your biology, but is to a significant extent within your own control.

Which one of the following is an assumption underlying the above argument?

A The latest study is more accurate than the previously accepted theories.

B The risk of developing cancer is simply down to extrinsic factors.

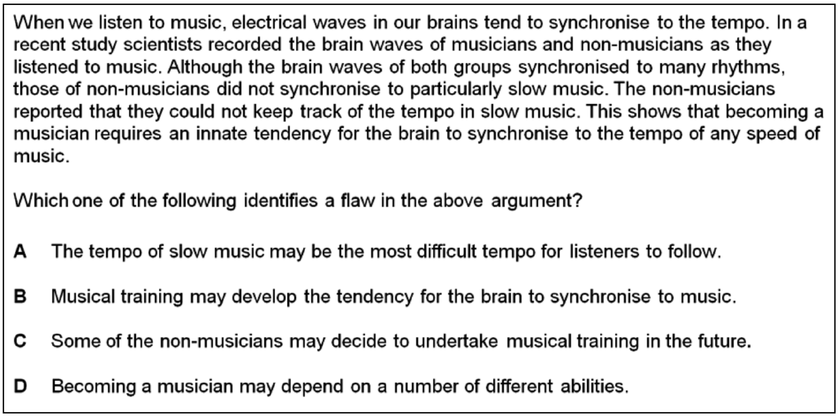
C There can be no ways of preventing the human body from making cancer-causing mistakes.

D People have some control over the influence of extrinsic factors such as stress or pollution.

presented in Figure 3.4. Again, an argument is presented in a short passage; however, this time, the question asks for the flaw in the argument.

Section 1 of BMAT and the critical thinking component in particular has been criticised in discussions on medical selection tests, which are important to address in any theoretical discussion of the test's construct. McManus et al (2005:557) argued that there is 'little agreement on what critical thinking means', and suggested that BMAT Section 1 is actually testing fluid intelligence. Whilst we concede that definitions of critical thinking are not

Figure 3.4 Example understanding argument (detecting flaws) item from BMAT 2016



When we listen to music, electrical waves in our brains tend to synchronise to the tempo. In a recent study scientists recorded the brain waves of musicians and non-musicians as they listened to music. Although the brain waves of both groups synchronised to many rhythms, those of non-musicians did not synchronise to particularly slow music. The non-musicians reported that they could not keep track of the tempo in slow music. This shows that becoming a musician requires an innate tendency for the brain to synchronise to the tempo of any speed of music.

Which one of the following identifies a flaw in the above argument?

A The tempo of slow music may be the most difficult tempo for listeners to follow.

B Musical training may develop the tendency for the brain to synchronise to music.

C Some of the non-musicians may decide to undertake musical training in the future.

D Becoming a musician may depend on a number of different abilities.

universally accepted, it is simply untrue that there is little agreement on the term’s meaning among educational assessment experts. Facione (1990) conducted a Delphi study and presented a statement of expert consensus on critical thinking that informed the work of Cambridge Assessment and many other critical thinking researchers. Although several disagree on the scope of critical thinking, with some researchers conceptualising a greater number of subskills than others (e.g. Paul and Elder 2007), there is agreement that Facione’s work captures the core elements of critical thinking ability, and that critical thinking skills can be developed through instruction (Halpern 1999, Sharples, Oxman, Mahtani, Chalmers, Oliver, Collins, Austvoll-Dahlgren and Hoffmann 2017). A number of assessments targeting critical thinking skills have been developed (Landrum and McCarthy 2015) and explored in higher education settings (O’Hare and McGuiness 2009). Furthermore, recent analysis using commercially available measures of critical thinking skills indicate that the construct is predictive of degree performance (O’Hare and McGuiness 2015).

The tendency by some to conceptualise BMAT as an intelligence test may be influenced by work on another admissions test used by medical schools, the United Kingdom Clinical Aptitude Test (UKCAT), which adopts a somewhat different approach to defining the construct to assess compared with BMAT (McManus, Dewberry, Nicholson and Dowell 2013). UKCAT was designed specifically to assess cognitive aptitude conceptualised as an innate construct that is intended to be independent from socio-economic factors. Therefore, UKCAT aims to assess an ‘innate ability to develop professional skills and competencies’ (Pearson VUE 2017:1), aligning the theoretical basis for the test with traditional IQ tests. However, critiques of BMAT describe and consider the constructs assessed by BMAT and UKCAT as interchangeable, particularly in relation to BMAT Section 1, based largely on assumptions about the degree to which BMAT scores might correlate with intelligence tests. From a cognitive validity perspective, we argue that the *a priori* approach advocated by Weir (2005) is important here, because there is little theoretical basis for treating the tests as identical.

Another critique of BMAT focuses on the fact that definitions of critical thinking often include a dispositional element. Specifically, these point out that 'critical thinking is related more to aspects of normal personality than it is to IQ' (McManus et al 2005:557). In our view, the idea that critical thinking includes dispositional aspects is not problematic. Facione (1990) clearly distinguishes between the skills and dispositional aspects of critical thinking, and BMAT Section 1 explicitly targets critical thinking skills, not dispositions, using the understanding argument items. Any comparison of BMAT Section 1 with measures designed to assess critical thinking disposition, such as Facione's (2000) California Critical Thinking Disposition Inventory (CCTDI) or Stupple, Maratos, Elander, Hunt, Cheung and Aubeeluck's (2017) Critical Thinking Toolkit (CriTT), would confirm BMAT Section 1's focus on skills.

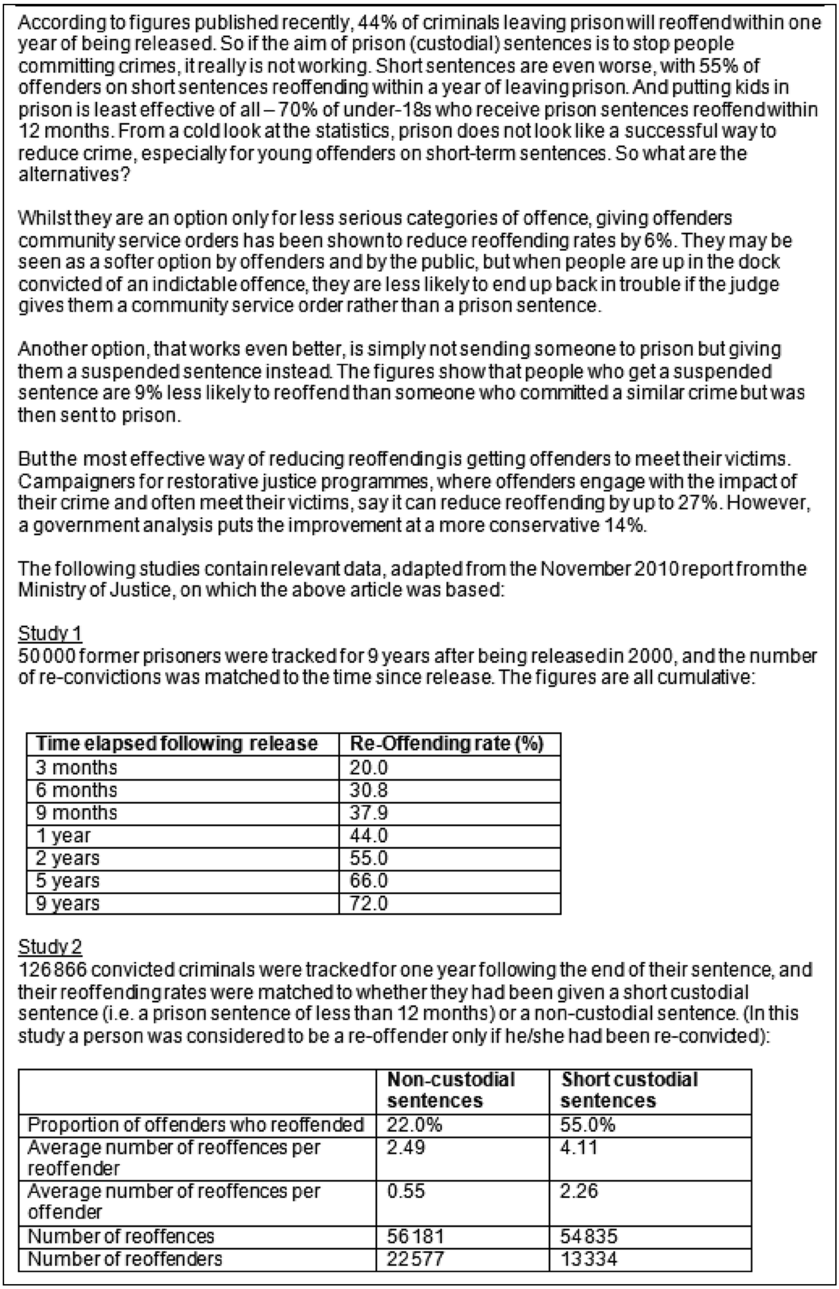
This misinterpretation of BMAT's test construct demonstrates the importance of articulating the cognitive processes that a test is intended to assess. Given that the rationale for assessing critical thinking skills is acknowledged in discussions about admissions tests, which is that 'critical thinking skills, not dispositions, predict success in examinations' (McManus et al 2005:557), it is possible that critiques of BMAT are not based on a full understanding of the cognitive processes assessed by the test. Bell et al's (2005) response to these criticisms clarified why BMAT should not be conceptualised as an intelligence test and the present chapter prevents further confusion by providing details on the theoretical underpinnings of BMAT's test sections, particularly on critical thinking.

Section 1 also includes data analysis and inference items, which require students to apply the skills described above to handling and interpreting larger amounts of information. These typically consist of sets of between three and five items associated with an extended passage of text, graphical and/or numerical data, and closely resemble a format used in Oxford's OMAT, which itself had adopted key features from the US Medical College Admission Test (MCAT) (James and Hawkins 2004).

Figure 3.5 presents the information provided for a set of four data analysis and inference items from BMAT 2015.

As shown here, a substantial amount of data is provided in a combination of forms, such as using written text and in tables. The amount of information can impact on how easy or difficult it is to sift through the content provided. As associated items are designed to target the candidate's data interpretation skills, the density of material presented is carefully monitored and adjusted during the item authoring process. The items associated with Figure 3.5 are available in Figure 3.6.

Figure 3.5 Example information provided for a set of data analysis and inference items from BMAT 2015



According to figures published recently, 44% of criminals leaving prison will reoffend within one year of being released. So if the aim of prison (custodial) sentences is to stop people committing crimes, it really is not working. Short sentences are even worse, with 55% of offenders on short sentences reoffending within a year of leaving prison. And putting kids in prison is least effective of all – 70% of under-18s who receive prison sentences reoffend within 12 months. From a cold look at the statistics, prison does not look like a successful way to reduce crime, especially for young offenders on short-term sentences. So what are the alternatives?

Whilst they are an option only for less serious categories of offence, giving offenders community service orders has been shown to reduce reoffending rates by 6%. They may be seen as a softer option by offenders and by the public, but when people are up in the dock convicted of an indictable offence, they are less likely to end up back in trouble if the judge gives them a community service order rather than a prison sentence.

Another option, that works even better, is simply not sending someone to prison but giving them a suspended sentence instead. The figures show that people who get a suspended sentence are 9% less likely to reoffend than someone who committed a similar crime but was then sent to prison.

But the most effective way of reducing reoffending is getting offenders to meet their victims. Campaigners for restorative justice programmes, where offenders engage with the impact of their crime and often meet their victims, say it can reduce reoffending by up to 27%. However, a government analysis puts the improvement at a more conservative 14%.

The following studies contain relevant data, adapted from the November 2010 report from the Ministry of Justice, on which the above article was based:

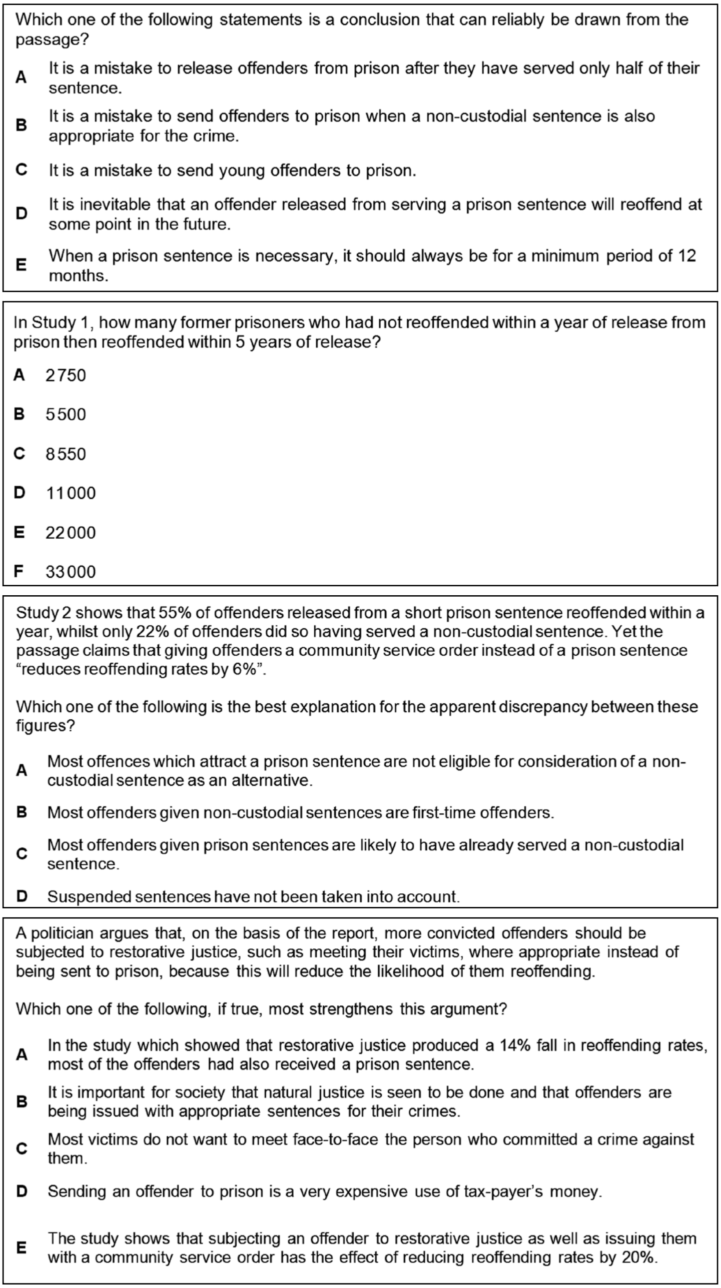
Study 1
50 000 former prisoners were tracked for 9 years after being released in 2000, and the number of re-convictions was matched to the time since release. The figures are all cumulative:

Time elapsed following release	Re-Offending rate (%)
3 months	20.0
6 months	30.8
9 months	37.9
1 year	44.0
2 years	55.0
5 years	66.0
9 years	72.0

Study 2
126 866 convicted criminals were tracked for one year following the end of their sentence, and their reoffending rates were matched to whether they had been given a short custodial sentence (i.e. a prison sentence of less than 12 months) or a non-custodial sentence. (In this study a person was considered to be a re-offender only if he/she had been re-convicted):

	Non-custodial sentences	Short custodial sentences
Proportion of offenders who reoffended	22.0%	55.0%
Average number of reoffences per reoffender	2.49	4.11
Average number of reoffences per offender	0.55	2.26
Number of reoffences	56 181	54 835
Number of reoffenders	22 577	13 334

Figure 3.6 An example set of data analysis and inference items from BMAT 2015



Which one of the following statements is a conclusion that can reliably be drawn from the passage?

A It is a mistake to release offenders from prison after they have served only half of their sentence.

B It is a mistake to send offenders to prison when a non-custodial sentence is also appropriate for the crime.

C It is a mistake to send young offenders to prison.

D It is inevitable that an offender released from serving a prison sentence will reoffend at some point in the future.

E When a prison sentence is necessary, it should always be for a minimum period of 12 months.

In Study 1, how many former prisoners who had not reoffended within a year of release from prison then reoffended within 5 years of release?

A 2750

B 5500

C 8550

D 11000

E 22000

F 33000

Study 2 shows that 55% of offenders released from a short prison sentence reoffended within a year, whilst only 22% of offenders did so having served a non-custodial sentence. Yet the passage claims that giving offenders a community service order instead of a prison sentence "reduces reoffending rates by 6%".

Which one of the following is the best explanation for the apparent discrepancy between these figures?

A Most offences which attract a prison sentence are not eligible for consideration of a non-custodial sentence as an alternative.

B Most offenders given non-custodial sentences are first-time offenders.

C Most offenders given prison sentences are likely to have already served a non-custodial sentence.

D Suspended sentences have not been taken into account.

A politician argues that, on the basis of the report, more convicted offenders should be subjected to restorative justice, such as meeting their victims, where appropriate instead of being sent to prison, because this will reduce the likelihood of them reoffending.

Which one of the following, if true, most strengthens this argument?

A In the study which showed that restorative justice produced a 14% fall in reoffending rates, most of the offenders had also received a prison sentence.

B It is important for society that natural justice is seen to be done and that offenders are being issued with appropriate sentences for their crimes.

C Most victims do not want to meet face-to-face the person who committed a crime against them.

D Sending an offender to prison is a very expensive use of tax-payer's money.

E The study shows that subjecting an offender to restorative justice as well as issuing them with a community service order has the effect of reducing reoffending rates by 20%.

These example items demonstrate how a set of data analysis and interpretation items include some overlap with problem solving and understanding argument items. However, candidates must also employ the skills that allow them to deal with larger quantities of information, in order to successfully complete data analysis and interpretation items.

Although the cognitive processes tested in Section 1 are seen as useful for learning across a range of subjects, they are also identified as particularly relevant for university courses in medicine and biomedical sciences. In the US, the Association of American Medical Colleges (AAMC) has identified 15 core competencies for entering medical students that are organised into four categories. Four competencies are listed in the thinking and reasoning category, including critical thinking and quantitative reasoning (Association of American Medical Colleges 2016); these competencies link to the skills assessed in BMAT Section 1. However, medical training is graduate entry in the US context; therefore it is important to consider medical education in the undergraduate setting.

In relation to the skills assessed by understanding argument items, the UK Quality Assurance Agency for Higher Education (QAA) benchmark statement for medicine courses states that 'graduates should demonstrate their ability to think critically by . . . adopting reflective and inquisitive attitudes and applying rational processes' (Quality Assurance Agency for Higher Education 2002:4). Problem solving is also referred to in the benchmark statement for medicine, but the statement for biomedical science provides the clearest link with the skills in Section 1 by identifying 'analytical, data interpretation and problem solving skills' (Quality Assurance Agency for Higher Education 2015:10) as attributes developed in a typical biomedical science course. The benchmark statements are intended to inform universities about the types of skills that various subject courses should develop, and this does not necessarily mean that the skills need to feature in university selection procedures; indeed, many attributes are listed in benchmark statements and only a subset are considered for admissions decisions. However, these links explicitly suggest that the skills assessed in Section 1 are relevant to successful undergraduate biomedical study, and contribute to the cognitive validity argument for assessing these processes.

Section 2 – Scientific Knowledge and Applications

The Scientific Knowledge and Applications section of BMAT adopted item types trialled in the Cambridge MVAT, for which there was evidence of a significant positive relationship between scores on the test and performance on the Cambridge Medical and Veterinary Science Tripos (Emery and Bell 2009). When deciding what might be assessed in Section 2, the key considerations were to develop a test of applicants' scientific understanding that would be accessible to candidates from a range of educational backgrounds, and would require candidates to do minimal additional new learning or preparation. Scientific knowledge is acknowledged as an important aspect of medical study, but there has been some debate about how best to assess science ability in the selection context. For example, McManus et al (2005:559) suggested commissioning a new test of 'high grade scientific knowledge and understanding' as a possible way of supporting medical selection. BMAT's Section 2 is a test of scientific knowledge and understanding, but it is unclear what would be needed to meet the criterion of 'high grade' for a test specification.

One important decision when developing BMAT Section 2 was that it should assess not only that a candidate has certain core scientific knowledge, but that they can apply it in a way that demonstrates an understanding of the scientific principles that underpin their knowledge, to distinguish a range of ability within a group of students with near-perfect grades or predicted grades in their school examinations. Therefore, Section 2 has an explicit focus on the cognitive processes involved in applying scientific knowledge to novel problems. This differentiates BMAT Section 2 from other science assessments administered as part of school qualifications, which typically include some questions testing recall of factual knowledge. Having access to a range of scientific knowledge is recognised as important for medical study; however, BMAT Section 2 is intended to complement school science qualifications rather than to serve as an alternative, hence the focus on applying school-level science knowledge to novel contexts.

There is some overlap between BMAT Section 2 and problem solving in Section 1; however, Section 2 items require problem solving skills to be applied to subject-specific knowledge. The current BMAT specification defines the skills and knowledge assessed by the BMAT Scientific Knowledge and Applications section as follows:

Box 3.4 The BMAT Section 2 test specification

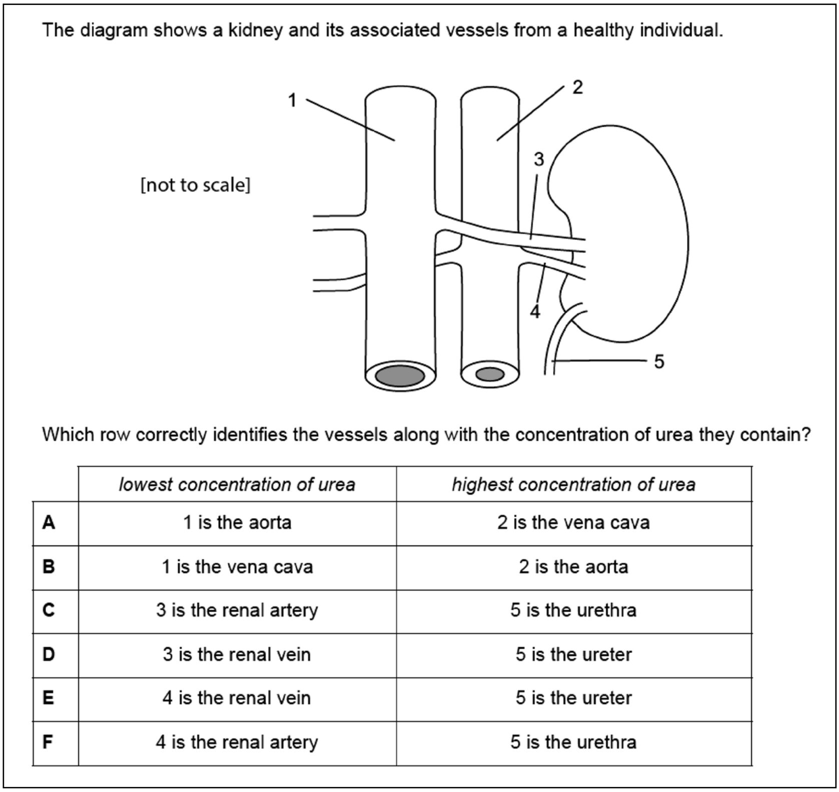
This element tests whether candidates have the core knowledge and the capacity to apply it which is a pre-requisite for high level study in biomedical sciences. Questions will be restricted to material typically included in non-specialist school Science and Mathematics courses. They will however require a level of understanding appropriate for such an able target group.

(Admissions Testing Service 2016b)

The design of BMAT Section 2 draws on a large body of research that conceptualises scientific thinking and reasoning as a form of problem solving (Dunbar and Fugelsang 2005). In this approach, scientific thinking is characterised as a search, or searches, in problem space (Simon and Newell 1971). BMAT Section 2 tasks can therefore be thought of as problems where the solution requires application of core science knowledge. This definition underpins guidance for item writers, who submit a description of the problem solution, which includes an account of the reasoning that a test taker needs to employ in order to solve the item. The solution for each item is analysed to make sure it involves a combination of both science knowledge and application. The nature of this analysis and the checks involved are described in Chapter 4.

A BMAT Section 2 item from 2016 is presented in Figure 3.7. This biology question requires a candidate to draw upon their knowledge of the anatomy of the kidney, their knowledge of the role of the kidney in the excretion of urea, and their knowledge of the structure and function of blood vessels. Candidates would not be able to answer this question by recall of this biological knowledge alone; instead, they must combine multiple aspects of their understanding to deduce the correct answer.

Figure 3.7 Example biology item from BMAT Section 2



The diagram shows a kidney and its associated vessels from a healthy individual.

1
2
3
4
5
[not to scale]

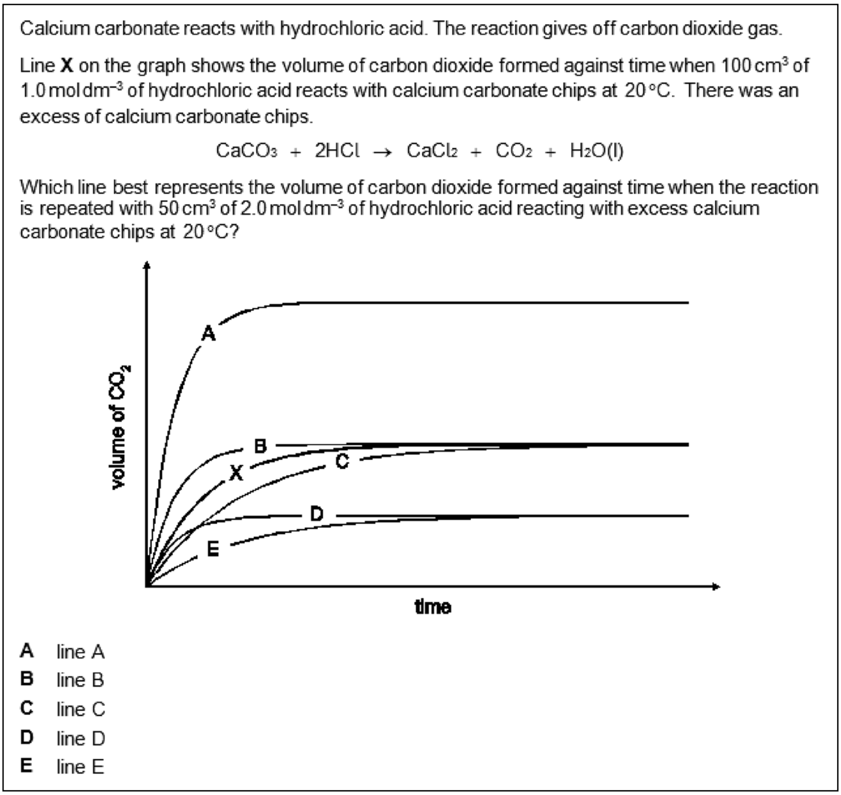
Which row correctly identifies the vessels along with the concentration of urea they contain?

	lowest concentration of urea	*highest concentration of urea*
A	1 is the aorta	2 is the vena cava
B	1 is the vena cava	2 is the aorta
C	3 is the renal artery	5 is the urethra
D	3 is the renal vein	5 is the ureter
E	4 is the renal vein	5 is the ureter
F	4 is the renal artery	5 is the urethra

Using their knowledge of the structure and function of blood vessels, a candidate should identify that vessel 2, which has thicker walls, carries blood away from the heart, and that vessel 1 carries blood returning to the heart. Using this information, and their knowledge of the anatomy of the kidney, they should be able to identify 3 as the renal vein and 4 as the renal artery.

A candidate should know that blood enters the kidney via the renal artery and leaves through the renal vein, and that a primary function of the kidney is to remove urea from the blood. This results in the production of urine, which leaves the kidney via the ureter, vessel 5. This information can then be used to deduce that the lowest concentration of urea will be in the renal vein (3) and that the highest concentration is in the ureter (5), so the correct answer is D. The incorrect options are based on candidates lacking knowledge of the function of the kidneys (A and B) or failing to correctly identify the vessels (C, E and F).

Figure 3.8 Example chemistry item from BMAT Section 2



Calcium carbonate reacts with hydrochloric acid. The reaction gives off carbon dioxide gas.

Line **X** on the graph shows the volume of carbon dioxide formed against time when 100 cm^3 of 1.0 mol dm^{-3} of hydrochloric acid reacts with calcium carbonate chips at 20 °C. There was an excess of calcium carbonate chips.

$$CaCO_3 + 2HCl \rightarrow CaCl_2 + CO_2 + H_2O(l)$$

Which line best represents the volume of carbon dioxide formed against time when the reaction is repeated with 50 cm^3 of 2.0 mol dm^{-3} of hydrochloric acid reacting with excess calcium carbonate chips at 20 °C?

A line A
B line B
C line C
D line D
E line E

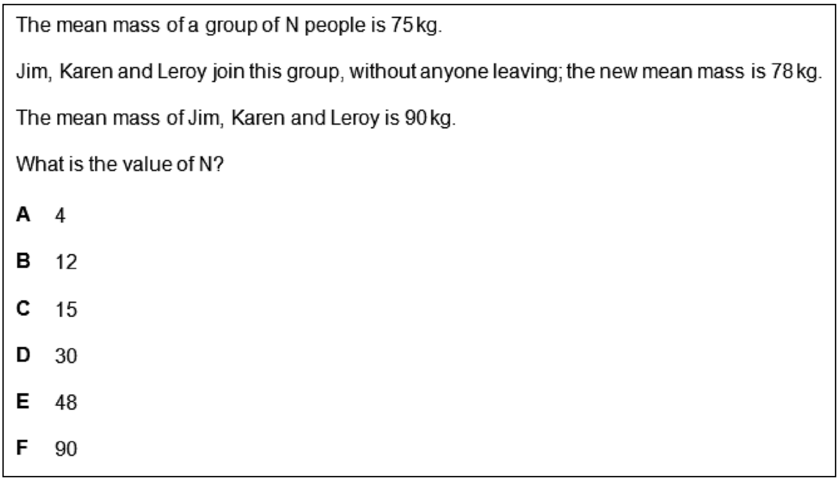
An example chemistry item is presented in Figure 3.8. The chemistry knowledge needed by the test taker includes understanding of a balanced chemical equation and familiarity with a formula for calculating the amount of substance (in moles) from a known volume and concentration of a solution. Knowledge of the specific reaction itself has not been assumed and the chemical equation is given to remove the need to construct it otherwise.

Using this subject-specific knowledge to successfully answer the question with B requires the test taker to recognise that as the calcium carbonate is in excess, the amount of carbon dioxide generated (proportional to volume measured at a given temperature and pressure) is only dependent on the amount of hydrochloric acid used. Using the given numbers, the amount of hydrochloric acid is the same in each experiment, and this observation narrows the choice of options down to line B or C. The gradient of line B is steeper than line X at the start of the second experiment, and the candidate needs to understand that in the second experiment, the initial rate of the reaction will be higher as the acid is more concentrated. The other response options are based on a misinterpretation of the gradients of the curves and/or mistaken use of the 2:1 ratio of hydrochloric acid to carbon dioxide in the chemical equation.

Another example of a BMAT Section 2 item is provided in Figure 3.9. This mathematics question requires the candidate to draw upon both their knowledge of how a mean is calculated and their facility with basic algebraic manipulation. The calculation of means is something students will have met early in their secondary mathematics education, and the algebraic manipulation needed in this question is straightforward. The candidate is required to assess the information given in the question and then to devise a strategy to move from that information to an answer. Any successful strategy adopted requires the candidate to possess a conceptual understanding of mean that reaches beyond rote learning; as such the question requires both the recall of elementary mathematical knowledge and the ability to assimilate given information whilst building a strategy that draws on a conceptual understanding of the relevant knowledge. The item's distractors are constructed to identify candidates who do not bring the conceptual understanding to their approach; for instance, E would appeal to candidates who fail to account for three extra people joining the group when they assimilate 78 into their strategy.

The physics item in Figure 3.10 requires the candidates to apply their knowledge of the relationship between mass, density and volume, alongside their facility with interpreting diagrams, extracting relevant information, and devising a strategy to solve the problem. Specifically, the test taker needs to know that density is equal to mass divided by volume; however, they must also have a conceptual understanding that the density of the material from which

Figure 3.9 Example mathematics item from BMAT Section 2



The mean mass of a group of N people is 75 kg.

Jim, Karen and Leroy join this group, without anyone leaving; the new mean mass is 78 kg.

The mean mass of Jim, Karen and Leroy is 90 kg.

What is the value of N?

A 4

B 12

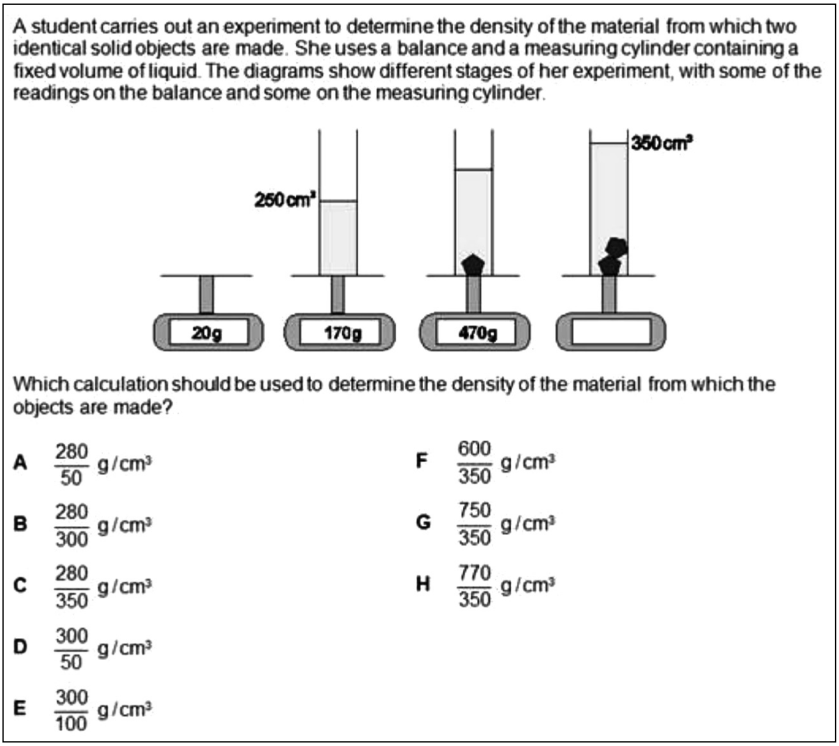
C 15

D 30

E 48

F 90

Figure 3.10 Example physics item from BMAT Section 2



A student carries out an experiment to determine the density of the material from which two identical solid objects are made. She uses a balance and a measuring cylinder containing a fixed volume of liquid. The diagrams show different stages of her experiment, with some of the readings on the balance and some on the measuring cylinder.

Which calculation should be used to determine the density of the material from which the objects are made?

A $\frac{280}{50}$ g/cm³

B $\frac{280}{300}$ g/cm³

C $\frac{280}{350}$ g/cm³

D $\frac{300}{50}$ g/cm³

E $\frac{300}{100}$ g/cm³

F $\frac{600}{350}$ g/cm³

G $\frac{750}{350}$ g/cm³

H $\frac{770}{350}$ g/cm³

the objects are made could be the mass of one object divided by the volume of one object or it could be the mass of two objects divided by the volume of two objects .

In order to solve the problem, these concepts must be applied to the information given, and the test taker must identify and ignore irrelevant information. The distractors are used to assess misconceptions in physics concepts or incorrect applications of skills and knowledge in devising a strategy. For example, distractor A would appeal to those who incorrectly use the initial reading on the balance to correct the final mass.

Including both knowledge and application in the specification of Section 2 recognises that scientific reasoning is sometimes understood as subject-specific conceptual knowledge and at other times as domain-general reasoning (Zimmerman 2000). Although we distinguish between task features that assess knowledge and those that are more related to reasoning, the approach employed for BMAT recognises that attempting to separate knowledge and strategy completely when operationalising scientific reasoning is highly artificial (Klahr and Dunbar 1988). For clarity, the knowledge and cognitive process elements of Section 2 are discussed separately in the present volume, as work defining the knowledge specification is described in Chapter 4. However, the theoretical perspective adopted for Section 2 acknowledges that they are often intertwined.

The UK General Medical Council's report on outcomes and standards for undergraduate medical education, *Tomorrow's Doctors* (2009), discusses the role of the 'doctor as a scientist' making explicit reference to the ability of the doctor to apply biomedical scientific principles and the scientific method to the practice of medicine. The AAMC (2014) also identifies scientific enquiry as a core thinking and reasoning competency for entering medical study in the US.

BMAT's Section 2 recognises the important role of scientific problem solving in undergraduate medical education. However, BMAT does not assess all of the scientific knowledge and reasoning needed to fulfil the role of 'doctor as scientist'; instead, the section is designed to address the needs of admissions tutors by ensuring that applicants' scientific understanding is adequate for the study of biomedical sciences, and that they can cope with the demands of a rigorous science-based course. BMAT Section 2 focuses on the application of basic science knowledge, which can be regarded as a pre-requisite for developing more advanced clinical reasoning. It is acknowledged that other perspectives on scientific reasoning are potentially important in the development of doctors, such as conceptualisation of scientific thinking as hypothesis formation and testing (Dunbar and Fugelsang 2005). The skills tested by BMAT Section 2 facilitate the development of further cognitive processes and skills, which are targeted in medical school and foundation-level training. Patel, Arocha and Zhang (2005:734) explain how basic science knowledge is inexorably linked to clinical problem solving (clinical reasoning):

> Basic science or biomedical knowledge is supposed to provide a scientific foundation for clinical reasoning. The conventional view is that basic science knowledge can be seamlessly integrated into clinical knowledge, analogous to the way that learning the rules of the road can contribute to one's mastery of driving a car.

The cognitive skills assessed by BMAT Section 2 focus on applying science knowledge to novel problems, as a precursor to extending this application to more complex decision-making scenarios specific to clinical contexts. Indeed the ability to apply science knowledge can be viewed as a set of skills that will improve with further training in clinical schools. This conceptualises them as abilities that can be developed, in line with the findings of research on scientific thinking skills and metacognition conducted with children (Zimmerman 2007, Zohar and Peled 2008). Furthermore, the ability to apply scientific knowledge to novel problems depends on use of cognitive and metacognitive strategies that are not assumed to routinely develop as part of childhood development (Morris, Croker, Masnick and Zimmerman 2012); indeed, the skills assessed in BMAT Section 2 are recognised as important skills to practise and develop.

Section 3 – Writing Task

It was noted in Cambridge Assessment's early work on admissions tests that while multiple-choice questions (MCQs) provided an objectively marked assessment which could address many of the elements of critical thinking, a test of students' 'productive' reasoning capacities – that is, the ability of candidates to produce a reasoned argument of their own – would also be welcomed by universities (Fisher 1992). These early observations of Cambridge Assessment researchers fit with more contemporary views on assessing higher order thinking skills expressed by international test providers, which advocate using multiple item formats in tests (Butler 2012, Ku 2009, Liu, Frankel and Roohr 2014). The recently revised MCAT dropped the writing task component based on limited evidence of predictive validity and an understanding that medical schools did not use the scores as extensively as other sections when ranking (Schwartzstein, Rosenfeld, Hilborn, Oyewole and Mitchell 2013). This decision was made despite the AAMC (2014) identifying written communication as a core thinking and reasoning competency for students entering medical study in the US. Removal of the writing task from MCAT prompted researchers at the Australian Council for Educational Research (ACER) to outline a cognitive validity argument for writing tasks that described how MCQs 'cannot reach into the cognitive recesses where a generative written task can' (McCurry and Chiavaroli 2013:570). Similarly, Cambridge Assessment's rationale for including a writing task in BMAT emphasises cognitive validity and the theoretical arguments for assessing written communication.

Communicating clearly in writing is a crucial skill that is sometimes underappreciated and regarded as secondary to clinical knowledge in medical contexts (Goodman and Edwards 2014). Oxford's OMAT test, as a precursor to BMAT, included a structured writing task based on the desire to examine 'communication through the use of clear written English to express abstractions and arguments' (James and Hawkins 2004:250). BMAT's Section 3 Writing Task requires candidates to produce a short piece of communicative writing on a topic of biomedical, general or scientific interest. It assesses the ability to select, develop and organise ideas, conveying them concisely and effectively.

The Writing Task is intended to complement the other BMAT sections, allowing demonstration of analytical reasoning skills and the ability to develop an argument, which extends the evaluation of these skills in a structured multiple-choice context in Section 1. In particular, the Section 3 task is designed to assess test takers' ability to construct clear and coherent arguments, which is an important part of Cambridge Assessment's definition of critical thinking (Black 2008). The current BMAT specification for Section 3 is in Box 3.5.

Box 3.5 The BMAT Section 3 test specification

Questions will provide a short proposition and may require candidates to:

- explain or discuss the proposition's implications
- suggest a counter proposition or argument
- suggest a (method for) resolution.

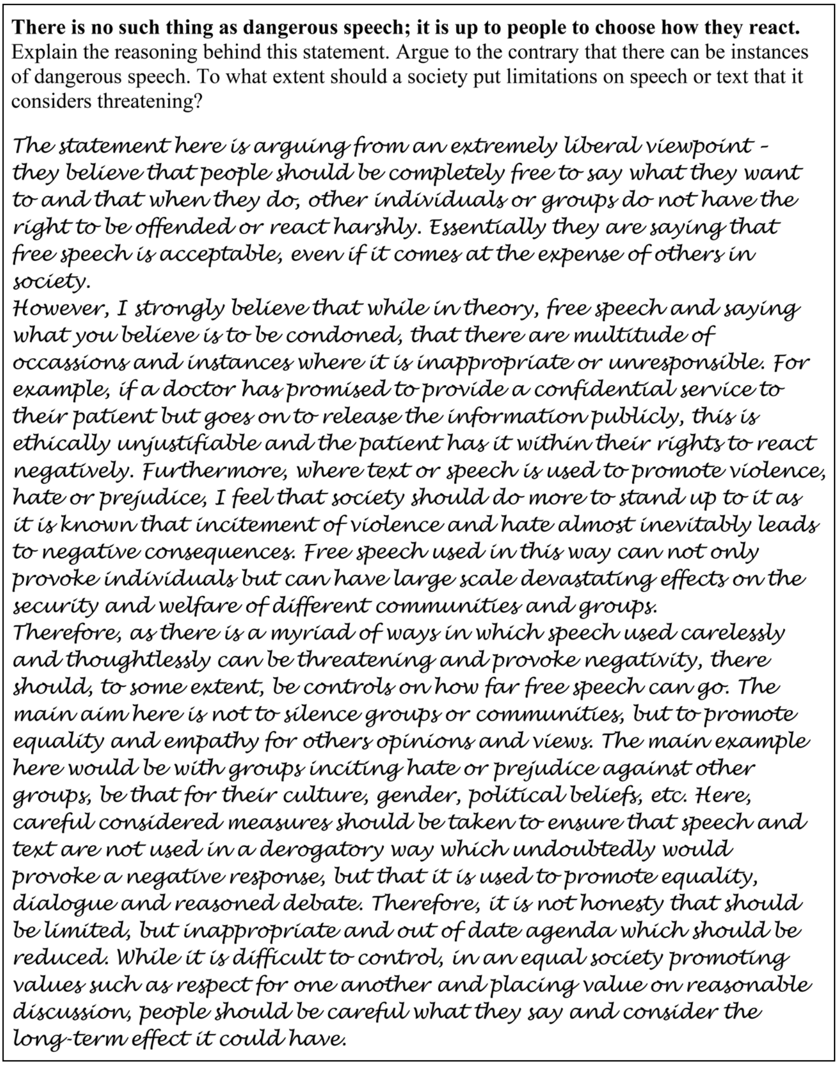
The Writing Task provides an opportunity for candidates to demonstrate the capacity to consider different aspects of a proposition, and to communicate them effectively in writing. Skills to be assessed include those concerning communication, described above. All specified skills may be assessed.

(Admissions Testing Service 2016b)

The approach to task design for Section 3 is informed by cognitive models of writing developed in psychology (Scardamalia and Bereiter 1987), which distinguish between knowledge-telling and knowledge-transforming strategies when writing. Knowledge-telling focuses on the topics and genres of a writing task to generate content. Knowledge-transforming, on the other hand, conceptualises a writing task as a rhetorical problem with goals and problems to overcome. Teaching English to Speakers of Other Languages

(TESOL) also target knowledge transforming in their writing components, but only at higher levels because it is an aspect of processing that can place great demands on non-native speakers of English. An example Section 3 question and response from BMAT 2014 is provided in Figure 3.11.

Figure 3.11 Example response to BMAT Section 3 that achieved a 4.5A



There is no such thing as dangerous speech; it is up to people to choose how they react. Explain the reasoning behind this statement. Argue to the contrary that there can be instances of dangerous speech. To what extent should a society put limitations on speech or text that it considers threatening?

The statement here is arguing from an extremely liberal viewpoint - they believe that people should be completely free to say what they want to and that when they do, other individuals or groups do not have the right to be offended or react harshly. Essentially they are saying that free speech is acceptable, even if it comes at the expense of others in society.

However, I strongly believe that while in theory, free speech and saying what you believe is to be condoned, that there are multitude of occassions and instances where it is inappropriate or unresponsible. For example, if a doctor has promised to provide a confidential service to their patient but goes on to release the information publicly, this is ethically unjustifiable and the patient has it within their rights to react negatively. Furthermore, where text or speech is used to promote violence, hate or prejudice, I feel that society should do more to stand up to it as it is known that incitement of violence and hate almost inevitably leads to negative consequences. Free speech used in this way can not only provoke individuals but can have large scale devastating effects on the security and welfare of different communities and groups.

Therefore, as there is a myriad of ways in which speech used carelessly and thoughtlessly can be threatening and provoke negativity, there should, to some extent, be controls on how far free speech can go. The main aim here is not to silence groups or communities, but to promote equality and empathy for others opinions and views. The main example here would be with groups inciting hate or prejudice against other groups, be that for their culture, gender, political beliefs, etc. Here, careful considered measures should be taken to ensure that speech and text are not used in a derogatory way which undoubtedly would provoke a negative response, but that it is used to promote equality, dialogue and reasoned debate. Therefore, it is not honesty that should be limited, but inappropriate and out of date agenda which should be reduced. While it is difficult to control, in an equal society promoting values such as respect for one another and placing value on reasonable discussion, people should be careful what they say and consider the long-term effect it could have.

The response in Figure 3.11 demonstrates knowledge transformation in the opening paragraph by expanding on the statement from the question to reach a broader conclusion. Importantly, the response does not merely describe the statement. Instead, the candidate infers arguments from the statement and extends the reasoning to comment on societal issues. The response also uses examples to develop and present a logical argument. As a result, it received a high score, as shown by the examiner's comments associated with this particular writing sample (see Box 3.6).

Box 3.6 Examiner comments for a BMAT Section 3 response that achieved a 4.5A

This response follows a clear plan; it is obviously structured by the components of the question. It begins with a clear definition and also explains the reasoning behind the statement (rather than just stating what it means), which shows that the writer has carefully read the question.

It uses a simple but relevant example to make its point and comes to a definite conclusion that society should monitor speech but not directly control it.

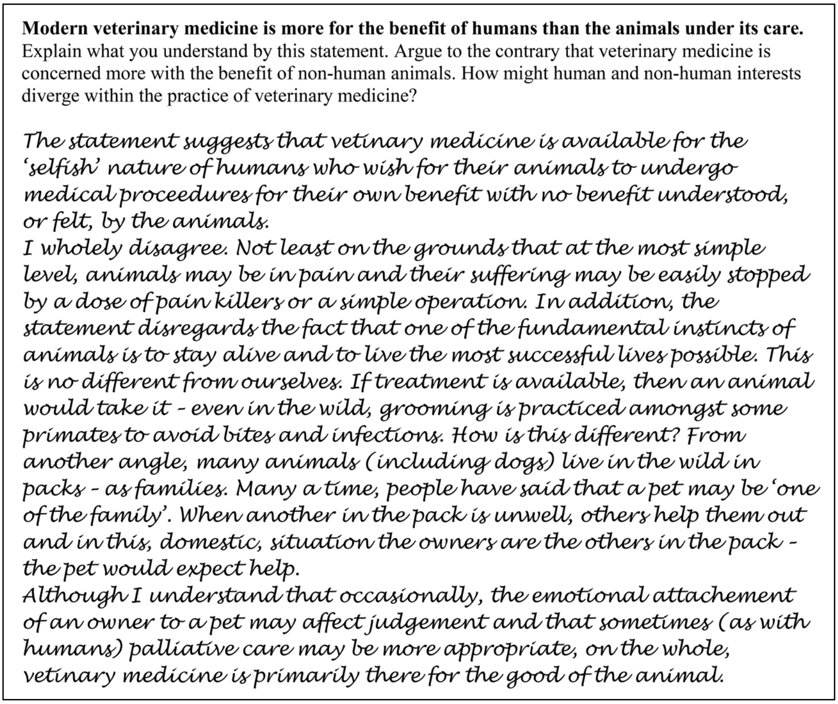
All aspects of the question are addressed effectively, providing a good counter-proposition. The argument is expressed in a clear and rational form, drawing things together into a balanced consideration of both sides.

Marks: 4.5A

The examiner comments for this response demonstrate BMAT Section 3's focus on argument and task completion. A response that exclusively used knowledge-telling strategies would not be able to achieve a high score, because knowledge transformation is needed to extend the opening statement from the question prompt (see Chapter 5 for details of the scoring criteria). According to Shaw and Weir (2007), careful task specification is needed to promote a writer's use of appropriate writing stages for English language tests. This observation also applies to the Writing Task in BMAT Section 3, which is structured to elicit the cognitive processes required to produce a well-structured argument; these processes include macro-planning, organising and monitoring one's text. The specific features of the BMAT Writing Task that support this are discussed in Chapter 4; however, it is clear from the examiner comments above that planning and organisation are needed for a Section 3 response to be scored highly. Another example Writing Task and response illustrates how a weaker answer tends to rely more heavily on knowledge-telling (Figure 3.12).

This response includes a number of relevant statements that are presented in isolation, but not joined together into a cohesive argument. Stating that many animals live in the wild is an example of knowledge-telling that does not extend the observation to make an overarching argument. The examiner comments (Box 3.7) associated with this response highlight that the counter-argument and conclusion fail to justify the response; they also omit any mention of planning or organisation.

Figure 3.12 Example response to BMAT Section 3 that achieved a 2A



Modern veterinary medicine is more for the benefit of humans than the animals under its care. Explain what you understand by this statement. Argue to the contrary that veterinary medicine is concerned more with the benefit of non-human animals. How might human and non-human interests diverge within the practice of veterinary medicine?

The statement suggests that vetinary medicine is available for the 'selfish' nature of humans who wish for their animals to undergo medical proceedures for their own benefit with no benefit understood, or felt, by the animals.

I wholely disagree. Not least on the grounds that at the most simple level, animals may be in pain and their suffering may be easily stopped by a dose of pain killers or a simple operation. In addition, the statement disregards the fact that one of the fundamental instincts of animals is to stay alive and to live the most successful lives possible. This is no different from ourselves. If treatment is available, then an animal would take it - even in the wild, grooming is practiced amongst some primates to avoid bites and infections. How is this different? From another angle, many animals (including dogs) live in the wild in packs - as families. Many a time, people have said that a pet may be 'one of the family'. When another in the pack is unwell, others help them out and in this, domestic, situation the owners are the others in the pack - the pet would expect help.

Although I understand that occasionally, the emotional attachement of an owner to a pet may affect judgement and that sometimes (as with humans) palliative care may be more appropriate, on the whole, vetinary medicine is primarily there for the good of the animal.

Box 3.7 Examiner comments for a BMAT Section 3 response that achieved a 2A

There is a simple and concise explanation of what is understood by the statement. The response clearly addresses all aspects of the question, although presenting the counter-argument as the candidate's own disagreement diminishes the force of the argument. The relief of an animal's pain could still be done purely for the benefit of humans. However, making the point that there is more similarity of interests between humans and animals than divergence is a good point. The counter-argument and conclusion are unconvincing – they fail to reasonably justify the response or to consider the whole of the argument around this topic.

Marks: 2A

Cambridge English writing exams that target knowledge transformation tend to include substantial stimulus materials for the writer to use, whereas BMAT Section 3 relies on a much shorter opening statement and the writer's own knowledge. This means that there are fewer opportunities to manipulate ideas when there is a limited pool of knowledge available to include in the response. Figure 3.13 shows a slightly stronger response to the same question answered in Figure 3.12. This example has a more

Figure 3.13 Example response to BMAT Section 3 that achieved a 3.5A

The relationship between humans and animals has evolved and changed over time. However, it is unequivocal that the current 'status-quo' of the owner-pet relationship results in a modern veterinary medicine which is tailored to the needs of the owner, more so than the pet. As humans, we have developed connections with animals that, in many ways, means that our well-being mentally, physically and psychologically is greatly impacted by the health of not only others of our kind, but also others who are non-human animals. Even in instances where the non-human animal is not a pet and not domesticated, our instincts are to further our knowledge by treating it and also gaining a sense of satisfaction, that we have been helpful.

Nonetheless, veterinary medicine involves animals, so it consequently benefits the non-human animals more. Treating animals ensures their survival or increases the longevity of their lives; therefore, the population of the species is more likely to survive, which is intrinsically more of a benefit for the animal than for humans. Animals are living beings as well, so to treat animals differently than what we would expect from health care for ourselves is unfair; the patient's needs must be the main priority.

On the surface, it appears that modern veterinary medicine has focussed more on humans and our expectations, but it impacts both animals and humans; to claim that only one is targetted would not be justifiable and certainly not something that can be judged or quantified. Both human and non-human interests diverge in veterinary medicine as the underlying basis of treatment is for the welfare of the animal, but the reasons are different. Humans own animals as pets for their pleasure and as a result of this, decisions made in relation to the animals will be based upon personal gain. If medicine only valued the well-being of animals, then treatment would be free, which would avoid the vast numbers of animals left untreated due to financial constraints. Although veterinary medicine is about animals, the reasons behind treatments vary.

sophisticated explanation of the statement from the prompt that extends beyond describing it. By developing points that build on each other in the opening paragraph, this response demonstrates knowledge transformation early on, but this opportunity could be easily missed. Therefore, it is possible that knowledge transformation is more difficult to achieve with Section 3 Writing Tasks when compared with writing papers designed specifically to assess language proficiency.

Although the opening argument of the response in Figure 3.13 is well structured, the rest of the response is not as organised. In addition, the examiner comments (Box 3.8) pointed out that some relevant areas were not considered, which may result from a lack of macro-planning. The example responses we have reviewed suggest that BMAT Section 3's Writing Task elicits use of knowledge transformation strategies by candidates. However, we should acknowledge that inferring activation of these strategies from reviews of written submissions is limited, due to the retrospective nature of this approach.

Box 3.8 Examiner comments for a BMAT Section 3 response that achieved a 3.5A

This has a strong start; it is well phrased and immediately engages with the candidate's understanding of the statement. It uses an interesting approach, taking the benefit to animals to be as a benefit to the species rather than to individual animals, but does not explain why veterinary care targeting only one of either animals or humans 'would not be justifiable and certainly not something that can be judged or quantified'. This is a good response, which is reasonably well argued but concentrates on pet owners with no consideration of livestock care or working animals. So it does not quite get into the marks for a good answer that makes effective use of material.

Marks: 3.5A

Compared with the writing components of Cambridge English's language examinations, BMAT Section 3 is a relatively short writing assessment. *Cambridge English: Advanced (CAE)* and *Cambridge English: Proficiency (CPE)* both include a Writing paper that is 1 hour and 30 minutes long, and each paper includes two separate tasks. The 30 minutes allowed for BMAT Section 3 might not provide the same opportunity for macro-planning and organisation that is given by longer tasks, because test takers may be more inclined to start writing without preparing a plan, despite the advice and instructions provided by Cambridge Assessment (see Chapter 4 for examples). However, it is important to acknowledge that many BMAT candidates are native speakers of English and all of them should have a high level of English language proficiency (see Chapter 2). This means that the majority of BMAT test takers are capable of engaging in knowledge transformation more quickly than typical test takers in a language testing context.

Writing samples, examiner comments and grade distributions indicate that the majority of candidates plan and organise their responses to Section 3. Nevertheless, it would be useful to directly investigate the degree to which macro-planning, organisation and monitoring are activated for individuals completing BMAT Section 3. Recent investigations of cognitive validity in language testing make use of verbal protocols (e.g. Bridges 2010), eye tracking (e.g. Yu, He and Isaacs 2017) and keystroke logging (Leijten and Van Waes 2013). So far, these techniques have not been used to investigate BMAT; employing these approaches with BMAT Section 3 could provide additional evidence of the test's cognitive validity.

Another area to investigate is the cognitive process of writing revision, which is not currently targeted by BMAT Section 3. The short time available for the Writing Task makes it unlikely that test takers have the opportunity to revise their responses. Further work could identify the impact that having more time for the written component of BMAT would have, because this potentially allows the task to assess candidates' abilities to revise their writing, which may fit with the applicant's potential to succeed at written tasks in biomedical study.

Although the cognitive validity of BMAT Section 3 draws on Cambridge English Language Assessment's work on writing assessment, the skills targeted by BMAT's Writing Task do differ slightly from those assessed in language tests. Notably, Section 3 focuses on the ability to organise and construct a cohesive argument. This means that aspects of critical thinking categorised as synthesis in Black's (2008) taxonomy (see Table 3.2) are being targeted. In particular, the skills targeted by BMAT Section 3 are the ability to construct a coherent, relevant argument or counter-argument and the ability to make and justify rational decisions.

Not only is it important to elicit the targeted cognitive skills, it is also crucial to reward successful use of these skills appropriately. Importantly, performance on a BMAT Section 3 task largely depends on the cogency and clarity of the argument in the response, so the criteria are aligned with skills from Black's (2008) taxonomy (see Chapter 5 for details of the grading criteria). Two scores are given to ensure that aspects relating to argument are given sufficient weight in the quality of content grade, which is considered separately to a quality of English grade. The quality of English grades tend to be negatively skewed, whereas the quality of content grades are more normally distributed, indicating that many, but not all, demonstrate the planning and organisation of ideas that are crucial for making strong arguments.

The use of two scores allows medical schools to primarily consider the quality of content score, unless the candidate has an unusually low quality of English score. Prior to 2010 a single score was awarded for Section 3, but institutions using BMAT requested that Cambridge Assessment distinguish a written response's argument quality from the quality of English demonstrated by the candidate. Assessment managers also noted that markers could be influenced by linguistic features and writing style in the response, which was not the main focus of BMAT Section 3. Scripts impacted by this were more likely to be identified for marking a third time by a senior examiner and the changes were made to respond to the ways that Section 3 was being used. The change in marking system made it easier for examiners to focus on construction of the argument when awarding the quality of content grade, which is the cognitive process that the task is designed to assess.

The first part of this chapter has focused on selecting and defining the cognitive processes that BMAT should assess, and the theories that underpin Cambridge Assessment's conceptualisation of these skills. The next portion of the chapter turns to research investigating how successful BMAT tasks are at targeting these skills.

3.4 Research on cognitive validity

Cognitive processes are difficult to investigate because they cannot be directly observed; instead their influence is inferred from indirect measures. In cognitive psychology, experimental methods have been used to uncover much of what is known about the way that individuals reason. A person's reaction time (Wilhelm and Oberauer 2006), where they are looking (Ball 2014) and even their brain activity (Goel, Navarrete, Noveck and Prado 2017) have been used to construct theories about the cognitive processes they are engaging. These methods and cognitive paradigms have been used to examine reasoning in a wide range of areas, such as problem solving (Gilhooly, Fioratou and Henretty 2010), hypothesis testing (Gale and Ball 2008) and deductive reasoning (Evans and Ball 2010).

Think-aloud studies

The kinds of data collected in experimental settings are rarely available in formal testing contexts, although developments in computer-based (CB) testing have encouraged some limited investigation with eye-tracking studies, particularly for tests of scientific problem solving (Tai, Loehr and Brigham 2006, Tsai, Hou, Lai, Liu and Yang 2012) and reading performance (Bax 2013). In educational assessment it is more common to collect data on cognitive processes by conducting think-aloud studies, which are sometimes referred to as 'cognitive labs', such as in the *Standards* (2014:82). Data from these studies is then interpreted using verbal protocol analysis, which is an established technique for gaining insight into the cognitive processes of candidates that would otherwise remain covert (Norris 1990). In this method, candidates (or research participants) are asked to 'think aloud', usually at the same time as they work through test items. The resulting information (the 'verbal protocol') is recorded, transcribed and then its content analysed according to a coding scheme.

Verbal protocols can be an account of how one *would* solve a problem, of how one *is currently* solving a problem or a retrospective account of how one *did* solve a problem. If performed concurrently with the task, candidates are asked to say out loud everything that goes through their head as they work through the task: to verbalise all their thoughts in the present tense. Retrospective reports from candidates of how they went about solving a problem have the advantage of lower interference with the task in hand but have the disadvantage of short-term memory decay. The nature of the test items may influence which method is the most suitable, for example, whether speed of processing is an aspect of the cognitive skill being assessed.

Early research by Cambridge Assessment on admissions tests for entry to higher education used verbal protocols to evaluate the cognitive validity of question types. The understanding argument and problem solving item types used in BMAT Section 1 have been investigated in this way with in-depth studies by Thomson and Fisher (1992) and Green (1992). These informed further development of thinking skills and reasoning tests including BMAT, and the findings were used to refine and improve the processes by which test questions are produced.

Green (1992) focused on questions similar to the problem solving items included in BMAT today. The analysis indicated that most items functioned well and that students did not approach them in a routine manner. Errors relating to routine execution and computational slips were low, suggesting that the items did not merely require the application of routine procedural methods. Green used a taxonomy by Mayer, Larkin and Kadane (1984) to plot phases of problem solving including understanding, method finding, planning and execution, and the associated knowledge that might be invoked such as linguistic and factual, schematic, strategic and algorithmic. Green's research overall found that the problem solving questions examined were appropriate but pointed to some areas for improvement. Notably, the findings highlighted two key features of items that can impact on the cognitive processes used by test takers:

- the language used in the item must be straightforward to ensure errors do not arise from problems with linguistic encoding, translation or understanding of the problem itself
- problems should not require reasoning that is counter-intuitive to real-life situations.

These findings informed guidelines for item writers which are used in authoring of BMAT questions today. In addition, reviews explicitly check linguistic features of items, and how the reasoning in each item relates to real-life situations. This ensures that construct-irrelevant variance is not introduced due to belief bias (Evans, Barston and Pollard 1983), which is the tendency to endorse arguments or solutions based on their believability in real-life settings rather than on their logical validity (Ball and Stupple 2016). The operational checks on BMAT items are designed to safeguard cognitive validity by checking the task features of items; some of these checks are described by Shannon, Crump and Wilson in Chapter 4 of this volume, which focuses on context validity.

Thomson and Fisher's (1992) research investigated questions similar to the understanding argument items in BMAT Section 1 using verbal protocol analysis. This work is presented in more detail below as a key study, to illustrate the research methods employed by Cambridge Assessment in these contexts.

Key study – A validation study of informal reasoning items (Thomson and Fisher 1992)

Main findings

- Think-aloud accounts of reasoning confirmed that test takers use targeted cognitive processes when answering the majority of tested MCQs.
- Complex wording and the design of options can result in items that assess reading comprehension rather than critical thinking.
- Minor changes and edits can reduce ambiguity and improve how well an item elicits the targeted skills.
- Terms used in MCQs assessing critical thinking are appropriate and understandable by test takers.

Introduction and context

Thomson and Fisher's (1992) study employed a similar approach to Green's (1992) investigation of formal reasoning items by using verbal protocols to examine informal reasoning items. This item type eventually became the critical thinking items in TSA and the understanding argument items in BMAT Section 1. The study was conducted as part of a larger pilot project (MENO) that trialled tests of six skills considered generally relevant to selection for university study (Willmott 2005). As part of evaluating the tests, the cognitive processes assessed by items were investigated using think-aloud studies. These were used to explore the theoretical underpinnings of the skills targeted using Cambridge Assessment's tests.

Research questions

The study explored types of items that now commonly appear in BMAT Section 1 and posed the following three questions:

- Do the items function as intended in the sense that candidates must reason correctly in order to answer the question correctly?
- What factors determine the difficulty of items (e.g. complexity of reasoning, language level, nature of distractors)?
- Do any items present obstacles which prevent candidates from demonstrating their reasoning ability? Do candidates understand terms such as 'main conclusion', 'assumption' etc.?

Data collection and analyses

Ten undergraduate participants (five each male and female, including two mature students) were interviewed individually, and after practice attempts, were asked to think aloud as they worked through 30 questions. Prompts to keep talking were given following long pauses. Most were able to tackle the questions and comment on their thinking as they progressed to a solution, although one participant found it particularly challenging. At the end, they were asked to reflect on what may have made particular questions difficult.

To facilitate coding of the transcripts, the test questions were analysed before the interviews took place to identify the reasoning intended to be necessary for correctly answering each item. The entire process, including the reflective interviews at the end, were transcribed verbatim and analysed by Cambridge Assessment researchers involved in the MENO thinking skills project.

Results

The percentage of correct answers given was calculated and participants' responses were categorised as 'reasoned well', 'no reasons given' or 'reasoned badly', by comparing the protocols with the reasoning process identified in analysis of the items. Instances where a correct answer was attained by poor reasoning were flagged for concern. Overall, the participants found the questions easier than expected, consistent with the fact they were undergraduates, rather than university applicants. Additionally, the standard time constraints were not imposed so more time was available to candidates to consider their answer.

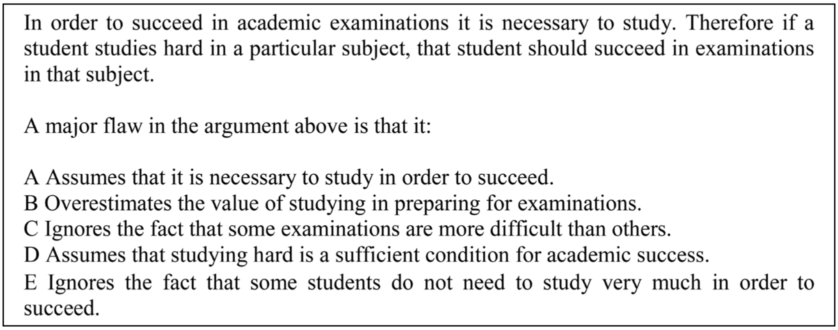
In total 20 questions were judged to work well and test the intended reasoning processes. It was suggested that a minor change to wording in three further questions would ensure they worked as intended – this was to clarify confusion by one or two participants only, so overall the questions were successful (see Figure 3.14 for an example). It was recommended that two items needed a replacement for one distractor, and two items were functioning badly and should be rejected, based on the complexity of the question text or weaknesses in the arguments or answer options.

This study provided valuable insight into the processes that candidates use when completing items similar to the understanding argument questions in BMAT Section 1. Several recommendations from this study were proposed and adopted for the development of future test items. The researchers recommended avoiding complicated wording, less common or technical words and convoluted sentences. They also flagged that some items appeared to be measuring reading comprehension rather than reasoning, and so recommended that where candidates are asked to identify main conclusions of a stimulus passage, the distractors should be components of the argument (e.g. reasons or intermediate conclusions); originally some questions had distractors that were not asserted in the stimulus passage, making it possible to answer the question by merely noticing that the correct answer is included in the passage whereas the other options were not, rather than through reasoning.

Overall, participants understood terminology related to critical thinking and could explain terms like ‘conclusion’, ‘assumption’, and ‘flaw in an argument’. Regarding a definition of conclusion, a number of participants referred to the ‘main message’ or ‘theme, idea, what it’s driving at’ in their response, which led to a recommendation that the main conclusion of an item needed to be the most interesting or focal point of the argument, rather than a more trivial but related point. Based on the participants’ verbalisations, all question subtypes functioned well and did not pose problems, provided that the stimulus passages and distractors were appropriately crafted.

The following example illustrates the in-depth process of analysing candidates’ verbalisations and using these to interpret the cognitive processes employed in responding to the question.

Figure 3.14 An example question analysed in the study



In order to succeed in academic examinations it is necessary to study. Therefore if a student studies hard in a particular subject, that student should succeed in examinations in that subject.

A major flaw in the argument above is that it:

A Assumes that it is necessary to study in order to succeed.
B Overestimates the value of studying in preparing for examinations.
C Ignores the fact that some examinations are more difficult than others.
D Assumes that studying hard is a sufficient condition for academic success.
E Ignores the fact that some students do not need to study very much in order to succeed.

In order to recognise D as the flaw in the argument, participants needed to see that the fact that something will not occur without a particular antecedent condition does not guarantee that this same something WILL occur if the antecedent condition is met. Those who gave the right answer reasoned as above ('The problem is that a student could study hard and still fail the exam') and their comments about the distractors did not indicate misunderstanding. One person rejected D because it talked about 'academic success' which he thought was more general than success in exams. Overall, six participants gave the right answer and reasoned well, while three reasoned poorly and gave an incorrect answer (choosing distractors A, C and E), and one candidate chose incorrect answer E but gave no reason. The results suggest the question is relatively difficult and that the distractors work well in tempting some candidates. However, the researchers also recommended that the wording of D should be amended for clarity.

Discussion

Cognitive validity studies such as Fisher and Thomson's (1992) demonstrate the complexity of developing questions that tap the relevant cognitive processes identified in test specifications. This research provided information to test developers about the structure and function of the types of questions that appear in BMAT Section 1, their capacity to measure the intended reasoning appropriately, and areas that could be targeted for improvement by question writers and editors. Findings from studies such as this inform the process through which questions are commissioned, reviewed and used in papers to ensure the best measurement performance of the questions, and the construct relevance of the test. The checks and reviews relevant to cognitive validity are outlined in a description of the question paper production process in Chapter 4.

Although useful for informing operational processes, there are specific drawbacks to these types of think-aloud studies and, as in the case of this particular research, the participants did not work under the same time constraints as in a real test; therefore their performances may be somewhat different from those elicited under exam conditions. In addition, there are some practical limitations to these studies that should be acknowledged. Data collection for think-aloud studies takes a large amount of participant time, so it is often only possible to conduct the study with a small number of participants, or a small number of items. There is also substantial time commitment needed on the part of the researcher. Whilst the richness of data captured by think-aloud studies is a strength, transcription and detailed analysis are painstaking processes that preclude these studies from being conducted regularly as part of operational processes. These issues mean that, whilst informative, generalisations from these kinds of studies need to be supported with other research into cognitive validity. One of the other approaches to cognitive validity is outlined in the following part of the chapter.

Latent constructs as cognitive processes

Some data from live test administrations is available to researchers interested in cognitive validity, such as candidate performances on each separate task or item. At an individual level, this information is not particularly useful; however, data from large cohorts can reveal if a test taker getting one item correct is more likely to get other particular questions correct. Statistical techniques known as factor analysis (FA) are used to examine performances on items and the relationships between them. These analyses can help a researcher understand whether the items in a test are all assessing one latent construct or whether subsets of items are testing separate constructs. Latent constructs are any variables that are not directly observable, which are often conceptualised as different cognitive skills or bodies of knowledge. In educational assessment, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are used to investigate the number of latent constructs being assessed by a test, which is commonly referred to as test dimensionality.

Investigating dimensionality has traditionally been considered as a problem for mathematical modelling that involves the application of various psychometric theories (McDonald 1981). EFA is used to indicate the number of latent constructs that can be theorised as present in the dataset. In addition to how many latent constructs might be present, the EFA identifies which items group together. It is then up to the researcher to interpret what is represented by each cluster of items *post-hoc*. CFA is a form of Structural Equation Modelling that is used when there are already ideas about which constructs might be observable in the dataset. In this approach the researcher specifies which items will group together to represent a latent construct or separate constructs. Essentially, CFA examines how well the data fits with previously determined models by examining the structure of item performances and the relationships between them. These statistical approaches to investigating cognitive processes are powerful tools, but Weir (2005:18) cautions against relying on these types of analysis too heavily:

> There is a need for validation at the *a priori* stage of test development. The more fully we are able to describe the construct we are attempting to measure at the *a priori* stage, the more meaningful might be the statistical approaches that can subsequently be applied to results of the test. Statistical data do not in themselves generate conceptual labels. We can never escape from the need to define what is being measured.

In light of these warnings, admissions tests developed by Cambridge Assessment are designed with an *a priori* definition of the skills and knowledge being assessed, and CFA studies are preferred over EFA ones, although they can be more technically challenging to conduct and interpret. A CFA study conducted by Cambridge Assessment researchers is presented as a key study here, after a brief note on dimensionality that highlights important issues to consider when investigating this aspect of a test.

A note on dimensionality

The dimensionality of a test relates to cognitive validity because dimensions identified using statistical methods can be interpreted as skills or cognitive abilities that are underlying an assessment. For example, data from a maths test might be analysed to show that it is assessing two latent constructs, which can be thought of as two aspects of mathematical ability, or two cognitive skills. The results indicate that the test is assessing two dimensions, but caution should be exercised when referring to a test as unidimensional or multidimensional. Like many other psychometric concepts, such as internal consistency, dimensionality is sometimes misleadingly attributed to a test. The statistical indicators used to make these claims actually reflect a dataset. Claims of dimensionality based on analyses are not strictly referring to a characteristic of the test *per se*, but rather to the performance of a particular group in a specific testing context.

This distinction is best illustrated by returning to our example. Consider the aforementioned maths test administered to primary school children. The results of the analysis indicate two dimensions and provide information on which questions, or items, relate to each dimension. All of the items for one dimension feature multiplication whereas those linked to the other dimension require addition. One can conclude that multiplication and addition are two separate abilities in the test. Now consider another administration of the same test to secondary school children who have had more maths teaching. In this cohort, you might expect those who have successfully mastered multiplication to have also learned addition, whereas low performers are likely to have general issues with their arithmetic skills. Analysis of this data is more likely to indicate that a single dimension is being assessed, even though the same test is being used.

Aside from the candidature, other aspects of the testing context might also impact analyses of dimensionality. For example, consider a result showing that the same test is assessing three dimensions, where the third dimension includes a mixture of multiplication and addition questions, but all of them require multiple steps. This finding could be difficult to interpret based on a review of the test in isolation, but an understanding of the test administration can be revealing. A low-stakes administration could explain the result if test takers did not complete questions requiring multiple steps due to the effort needed. This would suggest that the third dimension can be interpreted as motivation. Alternatively, if the test is administered under strict time limits, or if test takers were prohibited from writing down their calculations, this third dimension might be conceptualised as working memory capacity. In summary, the following points should be noted about dimensionality:

- dimensionality is not a property of the test alone
- claims about a test's dimensionality may not hold if the test is administered in different cohorts
- contextual factors such as motivation can impact on dimensionality
- most descriptions of test dimensionality are based on *post-hoc* statistical analyses of test sessions.

These issues should be considered when investigating dimensionality in the context of a test or test section. For brevity, the rest of the chapter refers to BMAT and BMAT sections as unidimensional or multidimensional, in line with conventions in the psychometric literature. However, the nuanced issues outlined here are considered by Cambridge Assessment researchers when conducting and reporting FA analyses, such as in the following key study.

Key study – Confirming the theoretical structure of BMAT using Structural Equation Modelling (Emery and Khalid 2013b)

Main findings

- It is valid to interpret Section 1 as measuring a unified construct of thinking skills.
- It is valid to interpret Section 2 as measuring a unified construct of scientific reasoning.
- There is some evidence that an aggregate score for BMAT Section 1 and 2 is appropriate.
- The MCQs in BMAT assess the intended constructs as defined in the test specifications.

Introduction

Cambridge Assessment Admissions Testing conducted the study described here to investigate the cognitive validity of BMAT through analysis of test performance data. Factor analysis was used to investigate the underlying factor structure of BMAT in its earliest years, following changes made to the structure of MVAT that resulted in the introduction of BMAT. More recently, research has been carried out to verify the theoretical structure of BMAT (Emery and Khalid 2013b) using CFA. Each of the three sections of BMAT theoretically measures a different construct, or set of cognitive skills, and each of these sections is assumed to be unidimensional. That is, each section is designed to measure a single construct. Candidates receive a single score for each BMAT section on this basis.

However, Sections 1 and 2 of the test each contain items belonging to various subtypes. BMAT Section 1 contains three item subtypes: problem solving, understanding argument, and data analysis and inference. BMAT Section 2 contains four item subtypes: biology, chemistry, physics and maths. We therefore wished to test the assumption that BMAT Section 1 and BMAT Section 2 are each unidimensional, rather than multidimensional, in nature.

Methods and models tested

As Section 3, the Writing Task, consists of a single item, only the dimensionality of Sections 1 and 2 were investigated. The item-level response data of a BMAT test cohort was analysed (BMAT 2011, N = 6,230 candidates). BMAT Sections 1 and 2 consist of 62 items in total: 35 items in Section 1 and 27 items in Section 2. LISREL software was used to conduct the CFA. Initial exploratory analyses (in SPSS Version 20) indicated that a single factor was suitable for all 62 test items. CFA models were therefore constructed on theoretical (i.e. test specification) grounds. For each section, a single-factor model (i.e. a model assuming unidimensionality) and a multi-factor model were specified. For Section 1, the multi-factor model tested was a three-factor model, with items specified as belonging to problem solving (13 items), understanding argument (10 items) or data analysis and inference factors (12 items). For Section 2, the multi-factor model tested was a four-factor model, with items specified as belonging to biology (seven items), chemistry (seven items), physics (seven items) or maths factors (six items).

Models were compared using five model-fit indices. Model-fit adequacy was judged against common reference values for these indices (Hu and Bentler 1999).

Results

For BMAT Section 1, model fit statistics were similar and indicated adequate fit for both the single-factor 'Aptitude and Skills' model and the three-factor 'problem solving, understanding argument, data analysis and inference' model. In the three-factor CFA model, the problem solving, understanding argument, data analysis and inference factors were highly correlated, supporting the notion that items in Section 1 are measuring a unidimensional construct. For BMAT Section 2, model fit statistics were similar and adequate for both the single-factor Scientific Knowledge and Applications model and the four-factor 'biology, chemistry, physics and maths' model. Again, the multidimensional four-factor model included strong correlations between the biology, chemistry, physics and maths factors, supporting the conceptualisation of Section 2 items as collective measures of a unidimensional construct. A final, two-factor model of all 62 BMAT items, with items specified as belonging to either Section 1 (35 items) or Section 2 (27 items), again showed adequate model fit.

These analyses provide evidence that the multiple-choice BMAT sections are assessing the intended cognitive skills. Weir (2005) cautions against over-interpreting *post-hoc* analyses of test scores as evidence of validity and presents FA studies as an example of this; however, the CFA approach for this study used *a priori* theorisations of the skills and their relationships with each other. Therefore, the results can be seen as confirmation of theories developed during test design, rather than a post-test model of the skills underlying BMAT. Note that the results do not completely rule out conceptualising each of BMAT sections 1 and 2 as multidimensional. However, using this alternative approach would potentially compromise scoring validity as shorter subtests would have low internal consistency (discussed in Chapter 5 of this volume). Combined with the theoretical basis for the two sections, the CFA provides evidence that Sections 1 and 2 are assessing two separate, but related, processes.

3.5 Chapter summary

This chapter has focused on the cognitive processes assessed by each section of BMAT. In accordance with Weir's (2005) original emphasis on theory-based validity as cognitive validity, the discussion has included relevant models from educational assessment, science education and writing assessment. These have informed the theoretical basis for assessing thinking skills, scientific problem solving and written communication in BMAT. Additionally, some of the difficulties facing researchers interested in cognitive validity were outlined. While acknowledging the limitations in conducting research on cognitive validity, Cambridge Assessment Admissions Testing has conducted significant research in this area and two examples are presented in the chapter as key studies. These represent some of the more common approaches to investigating cognitive processing in educational assessment. The descriptions of these studies are potentially useful for researchers who are interested in the concept of cognitive validity, but are unfamiliar with methods used more widely in educational assessment. It is hoped that this encourages consideration of cognitive validity, both in the design and the evaluation of assessments, particularly in smaller scale contexts where bespoke tests or methods are being used.

On the other hand, considering relevant educational and psychological theory is no doubt a familiar practice for the seasoned test developer. Despite it being a core issue in educational assessment, test providers do not always describe the theoretical bases for their assessments, possibly under the impression that few individuals outside of testing fields will be interested. We see the presentation of theory underlying a test as an important responsibility of the test provider, because it allows the cognitive processes targeted by an assessment to be interrogated and challenged, particularly by users of the test. In the case of BMAT, it is important that medical schools are able to evaluate the reasons for including the constructs present in the test, even if only to support their communication with prospective applicants.

Although the discussion of cognitive processes assessed by BMAT in the current chapter has been detailed in comparison to the approaches adopted for some other assessments, they are still not as extensive as they might be. In particular, our understanding of the interaction between subject-specific knowledge and domain-general reasoning could be investigated further for BMAT Section 2 items. Also, the possible limitations of BMAT Section 3 should be considered in light of its short length and reliance on a single task.

Eye-tracking, key-logging and CB-testing technologies present greater opportunities for investigating these areas and others. They can complement traditional methods such as think-aloud studies, to potentially further our understanding of the cognitive processes elicited by educational assessments, which can inform the theory and practice of assessing all kinds of learning. In addition, investigating these issues in educational assessment could improve models and frameworks used in other fields, such as cognitive psychology. Collaboration between psychologists, medical educators and assessment experts is likely to support these endeavours, and multidisciplinary approaches should be encouraged in research and practice. Perhaps this recommendation is unsurprising, given the authors contributing to the present volume; however, we should point out that other disciplines have plenty to offer. Early indications and collaborations suggest that machine learning could have paradigm-changing impacts on our understandings and models of educational assessment.

The role of theory in informing assessment is not limited to future development. The opportunities afforded by cross-disciplinary collaboration and technologies are potential, whereas current test construction practices are actively informed by understandings of theory. Weir (2005) recognised that the interaction between theory-based and context-related aspects of validity is crucially important when considering overall construct validity. The present chapter has touched upon the relationship of cognitive validity with context-related validity, and with scoring validity. In the following chapters, these relationships are explored in greater detail, starting with the ways that context validity is informed by cognitive theories underlying an assessment. The next chapter details how decisions taken on item design and task setting affect the cognitive processing required to successfully complete test items in BMAT.

Chapter 3 main points

- Explicitly identifying the cognitive processes targeted by a test such as BMAT allows linking to relevant theories.
- Cognitive validity has been investigated in BMAT using verbal analysis protocols and factor analysis studies.
- Eye-tracking and computer-based testing present other opportunities for better understanding cognitive validity, particularly for Sections 2 and 3.
- Understanding the theoretical basis for any assessment can improve the design and production processes.

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.