Cambridge Assessment

Research *Matters*

Issue 29 / *Spring 2020*

Cambridge Assessment

Proud to be part of the University of Cambridge

Established over 150 years ago, Cambridge Assessment operates and manages the University's three exam boards and carries out leading-edge academic and operational research on assessment in education. We are a not-for-profit organisation.

Citation

Articles in this publication should be cited using the following example for article 1: Crisp, V., & Macinska, S. (2020). Accessibility in GCSE Science exams – Students' perspectives. *Research Matters: A Cambridge Assessment publication, 29*, 2–10.

Credits

Reviewers: Matthew Carroll, Gill Elliott, Tim Gill, Sylvia Vitello, Joanna Williamson, Research Division, Cambridge Assessment
Editorial and production management: Anouk Peigne, Research Division, Cambridge Assessment
Additional proofreading: Hannah Williams, Research Division, Cambridge Assessment
Cover image: John Foxx Images
Design: George Hammond
Print management: Canon Business Services

All details are correct at the time of going to print in April 2020

If you would like to comment on any of the articles in this issue, please contact Tom Bramley, Director, Research Division. Email: researchprogrammes@cambridgeassessment.org.uk

This and all previous issues of *Research Matters* are available to download, in full and by individual article, from our website: www.cambridgeassessment.org.uk/research-matters

Research *Matters* / 29

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

'Improved accessibility' has been vigorously pursued in contemporary assessments in England, and has featured in other many nations' discussions of fairness and bias. Perhaps it would better be described as 'removal of sources of construct-irrelevant score variation'. This better describes the relation between item quality and item purpose. Increasing assessment quality for those with sight or other impairment is essential where irrelevant features of items (font size, font type, colour, etc.) can readily be adjusted to improve the measurement properties of items and assessments. But some work on accessibility can impact adversely on the measurement properties of items and assessments. For example, Isaac Physics – its leading developers recently rewarded with a prestigious gold medal at the 2019 Institute of Physics awards – has highlighted how some efforts to improve accessibility (providing a diagram where none was expected before; breaking a question down into steps; providing equations) can materially impact on standards of demand, with negative washback into learning. 'Improving accessibility' is not some discrete and pre-eminent concern in assessment, since pursuing some accessibility aims can have a very specific, adverse impact on standards of demand. As is so often the case in assessment, complex things are entwined in complex ways. The best policy scenario is that the tension between enhanced accessibility and maintenance of standards is held in careful balance. The worst scenario is that the tensions lead to oscillations in priorities, and the resulting pendulum swings create precisely the kind of constant change in qualifications which educational professionals and candidates find disruptive and frustrating. The tensions will never go away; but sound and careful management means that adverse effects can be minimised.

Tim Oates, CBE *Group Director, Assessment Research and Development*

Editorial

Writing good exam questions is a difficult art. We need questions that elicit responses that demonstrate the relevant knowledge, skills and understanding (KSU). We want to avoid anything that hinders or prevents examinees with the relevant KSU from demonstrating this, so we should make the questions as accessible as possible without reducing their subject-related demands. The first two articles in this issue are about accessibility. The first, by Victoria Crisp and Sylwia Macinska, describes students' reactions to 'before and after' versions of exam questions that had various features modified in ways that were hypothesised to affect accessibility. The second, by David Beauchamp and Filio Constantinou, explores the potential of automated tools to give insights about the linguistic complexity of the words and sentences in exam questions.

In the third article, Stuart Shaw, Victoria Crisp and Sarah Hughes describe a rigorous but practical approach that could help practitioners to investigate the comparability of alternative assessments. Their framework distinguishes different kinds of standard and helpfully recognises that an overall informed judgement is required about the extent to which differences in comparability matter.

The final two articles, by me and Tom Benton respectively, explore an issue that is of perennial interest to assessment developers – namely the extent to which expert judgement about the difficulty of exam questions can give useful information about the relative difficulty of two exams as experienced by the examinees. The conclusions are somewhat pessimistic, but no doubt this will not be the last word on this topic!

Tom Bramley *Director, Research Division*

Accessibility in GCSE Science exams – Students' perspectives

Victoria Crisp and **Sylwia Macinska** Research Division

Introduction

The main purpose of many educational assessments is to measure students' achievement in relation to the construct(s) of interest. Therefore, any differences in students' outcomes should be due to the ability of the students with respect to the relevant construct(s). Students' performance on the test, however, is often a result of the interaction between multiple factors in addition to students' ability (Beddow, Elliott, & Kettler, 2013; Crisp, 2011; Spalding, 2009). These factors can relate to intrinsic student characteristics (e.g., test anxiety or working memory capacity) or to the construction of the test itself.

There are multiple elements of question design that can influence a student's ability to understand the question and demonstrate their achievement. These may include (but are not limited to) visual features, such as the use of images, legibility (font), layout of the question and linguistic complexity. If the questions present accessibility problems, then the resultant performance on the test may not reflect the students' achievement in relation to the construct(s), but rather their ability to access the meaning of the question (Beddow, Kurz, & Frey, 2011). Research shows that different elements of question construction can affect students' perceptions of accessibility and/or students' performance (Chelesnik, 2009; Crisp, 2011; Crisp & Sweiry, 2006; Lonsdale, Dyson, & Reynolds, 2006). Even small changes to question presentation, such as highlighting a key word using bold font style, can potentially lead to increased student success on the question (Pollitt, Ahmed, & Crisp, 2007). The aim of improving the accessibility of a question is not to reduce its demands but to provide students with a better opportunity to demonstrate their knowledge and skills by removing any obstacles to question comprehension. By demands we mean the knowledge and skills that will be needed in order to complete a task and that have been intentionally included in a question (Pollitt et al., 2007). These demands, which relate to the assessment constructs, are expected to determine how difficult a task is in practice, but other factors (such as question features that influence accessibility) can also affect difficulty. Optimising features in terms of accessibility allows students to better show their abilities related to the target construct(s) by keeping construct-irrelevant variance to a minimum (Ahmed & Pollitt, 2011).

The design of the question has the potential to either minimise or emphasise differences between students' characteristics. Accessibility-related features of the question interact with the intrinsic characteristics of the test taker such as motivation, reading comprehension and working memory capacity (Beddow et al., 2011). Changes to accessibility may therefore indirectly affect students' outcomes, even if the construct-related demand of the question remains the same. For example, embedding a question in a complex context risks introducing linguistic bias, therefore emphasising reading comprehension differences between students (Ketterlin-Geller, 2008). Similarly, text presentation that maximises the use of 'whitespace' (i.e., the part of the page not covered by text or images) influences how friendly or intimidating the text is perceived to be (Baker, 2001), which may affect students' motivation or test anxiety.

Students may find it frustrating if they are not able to understand the question, especially if they have mastered the construct that is being examined. If the test is perceived as difficult, students' experience of sitting the test is likely to be negative, regardless of the actual outcomes. Therefore, it is important to determine how different question features contribute to the perception of accessibility in the target assessment population.

Research context and aims

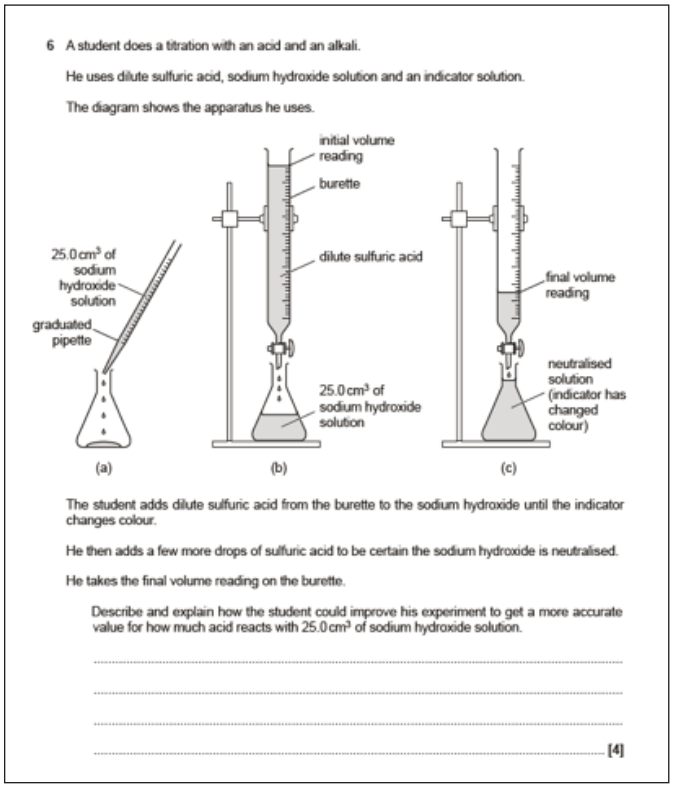
For some time, there has been a regulatory requirement for awarding bodies in England to "consider the needs of all potential candidates when developing qualifications, associated tasks and assessment, to minimise any later need to make reasonable adjustments for candidates who have particular requirements" (QCA, 2004, p.12). This is part of a notion of incorporating fair access for all students into assessment design (QCA, 2005). OCR has recently developed accessibility principles for Science GCSE exams (OCR, 2018a; 2018b), which intend to facilitate improvements to question design that enable students to show their knowledge and skills to the best of their ability. The principles draw on past research on the effects of question features on test accessibility. OCR first applied the accessibility principles when developing the GCSE Science question papers sat in the June 2018 session, as part of a question paper review process before the final sign off. The principles have also been applied to the sample assessment materials and practice papers.

The aim of the current research was to evaluate the effectiveness of OCR's accessibility principles by investigating students' perceptions of question features in terms of accessibility. Specifically, the research sought to determine whether question features relating to the accessibility principles affect students' views on how easy questions are to understand. To this end, we used a selection of Science GCSE exam questions, with and without the accessibility principles applied, to gather student views on relevant question features.

Method

Selection of questions

For the purpose of this research, OCR provided six Foundation tier Science GCSE papers from the June 2018 session. There were two versions of each paper: the final version of the paper as used in the live examination (with accessibility principles applied); and the draft of the



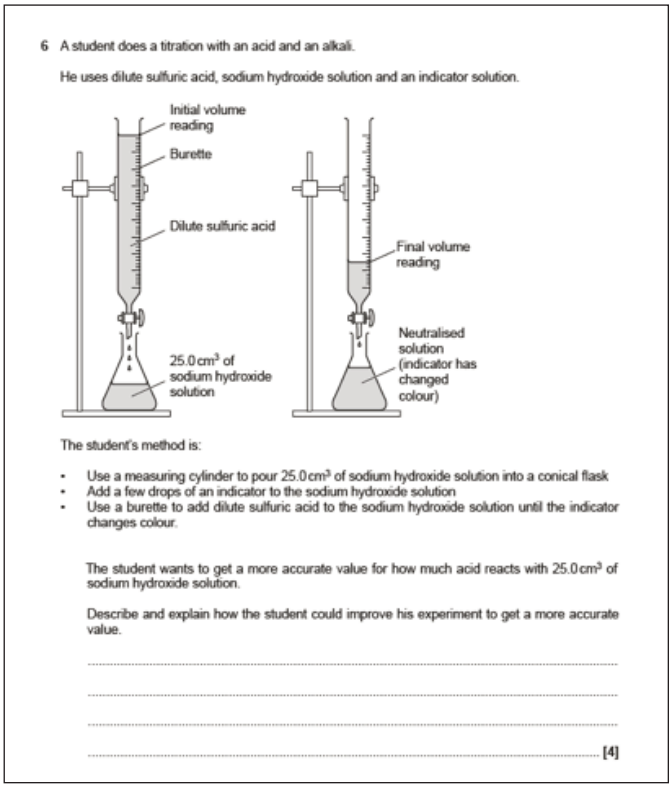


Figure 1: Two versions of an example question used in the test. Left panel: draft question before the accessibility principles were applied. Right panel: the final version of the question (after the accessibility principles were applied).

paper before the accessibility principles were applied. We compared the two versions of the papers, identifying questions where the changes were clearly due to, or aligned with, the accessibility principles. From this, we selected eight questions that were then renumbered as Questions 1 to 8.

The eight questions were included in both versions of a test. Version 1 of the test contained the final versions of Questions 1, 3, 5 and 7 (with the accessibility principles applied) and the draft versions of Questions 2, 4, 6 and 8 (without the accessibility principles applied). Version 2 of the test contained the opposite pattern. In this article, we refer to the question versions without the accessibility principles applied as 'less accessible' (LA) and the versions with the accessibility principles applied as 'more accessible' (MA), though it should be noted that these labels reflect the intentions to improve accessibility and may not always match student views. Figure 1 shows the two versions of an example question (Question 6) used in the research. Both versions of each question are available in an appendix to the online copy of this article.

The questions covered a range of the accessibility principles. Table 1 presents the accessibility themes explored, their relationship to OCR's accessibility principles and which question(s) were used to explore each theme. OCR's accessibility principles are reproduced in an appendix to the online copy of this article.

Participants and procedure

Four schools participated in the research (two comprehensive, one independent and one independent special provision), with one or two Year 11 Science classes taking part at each school. All students in participating classes completed one version of the test, with the two versions of the test assigned at random within each class. We interviewed 57 students across the schools after they had taken the test. The teachers selected students so that we could cover a range of abilities. Students had the opportunity to decline. In most cases, we interviewed students in pairs, where each pair included one student who took each version of the test. We discussed each question in turn, encouraging students to talk about how accessible the questions were and why, and gathered comparative comments in relation to specific accessibility-related differences between question versions. To help students understand the notion of accessibility we used wording such as 'easier to understand'. Where students' responses suggested that they might be commenting about question demands rather than accessibility, further prompting was used to gain responses relating to accessibility.

Results

Findings for each test question

We categorised students' responses regarding whether they understood the version of the question that they attempted as 'yes', 'no' or 'unclear/mixed' (no explicit comment or mixed opinion).

We categorised comparative views regarding each relevant accessibility theme as:

- V1 (Version 1 considered easier to understand than Version 2);
- V2 (Version 2 considered easier to understand than Version 1);
- no difference (no difference in perceived ease of understanding between versions);
- unclear/mixed (no explicit response/mixed opinion).

The findings for each question are now described in Tables 2 to 9 which show the results for each question. Percentages are used for ease of interpretation, but it should be noted that these are based on relatively low numbers: 28 students who attempted Version 1 of the test (V1);

Table 1: Accessibility themes explored, their relationship to OCR's accessibility principles and the question(s) used to explore each theme

Accessibility theme		*Relevant accessibility principle (OCR, 2018a, pp.5–7)*	*Biology*	*Chemistry*	*Physics*
Language:	– Simplified vocabulary	Principle 2			Q3
	– Simplified grammatical structure				Q7
	– Clarity of information			Q6	Q3
Presentation of context:	– Shorter context	Principle 4[1]	Q2, Q4		
	– Use of bullet points		Q8	Q6	
Multiple choice question (MCQ) answers in alphabetical order/numerical order		Principle 8		Q1	Q7b
Brackets used around abbreviations for units		Principle 10			Q7b
Visual resources:	– Only use where necessary	Principle 13[2]	Q2	Q6	Q5
	– Clarity of visuals				Q3
Left-aligned (tables/graphs)		Principle 14	Q8		
		Total number of questions:	**3**	**2**	**3**

1. This principle does not explicitly mention shortening a context, but the need for supportive devices such as bullet points in longer contexts implies that a shorter context (or no context) may have benefits for accessibility. There is some evidence that word count can influence student performance, for example, OECD (2009) found that word count accounted for 12% of variance in question difficulty, which could be due to reading demand affecting accessibility.
2. The clarity of visual resources is not explicitly stated as an accessibility principle but is likely to be important (Crisp & Sweiry, 2006).

29 students who attempted Version 2 of the test (V2); and 57 students in total. Therefore, care should be taken not to over-interpret differences. Note that percentages have been rounded to whole numbers, which has sometimes resulted in values that add up to over 100%.

Students' comments provided insights into the reasons for their views. Common explanations for their views about accessibility are included below.

Question 1

Question 1 was a multiple choice question asking students which statement about catalysts was correct. It was selected to investigate whether the order of answer options influenced students' perceptions of ease of understanding. Answer options appeared in alphabetical order in one version of the question (more accessible version) and in a random order in the other. Over 80% of students found Question 1 easy to understand, regardless of which version they had attempted. When asked to compare the question versions, the majority of students (84%) reported that the order of the options made no difference to the ease of understanding and answering the question. The most common comments justifying their position were that they would be able to select the correct answer regardless of the order, as long as they had the appropriate knowledge, and that they would read all options anyway.

Table 2: Frequencies of responses regarding Question 1 (Catalysts)

Was the question easy to understand?	**V1 More accessible (MA)**	**V2 Less accessible (LA)**		
Yes	23 (82%)	26 (90%)		
No	2 (7%)	2 (7%)		
Unclear/mixed	3 (11%)	1 (3%)		
Order – which is easier to understand?	**V1 – MA** (alphabetical order)	**V2 – LA** (random order)	**No difference**	**Unclear/ mixed**
Frequency	3 (5%)	5 (9%)	48 (84%)	1 (2%)

Question 2

Question 2 was selected to explore the influence of context and visuals on accessibility. The question required students to categorise four human characteristics as either continuous or discontinuous. The less accessible version of the question included a context about two sisters, information on some of their characteristics (e.g., 'Height = 150 cm') and cartoon-style images; both the contextual information and the images were removed in the more accessible version. For both versions, most students reported that they understood the question.

When asked to compare the question versions in terms of context use, the contextualised version was more frequently perceived as harder to understand than the context-free version (the latter was preferred by 58% of students). Students typically reported that they liked the clear presentation of the list of characteristics in the more accessible version. Some students were confused by the examples of characteristics in the less accessible version and felt it was unclear whether to report the characteristics themselves (e.g., 'Height') or the examples provided (e.g., '150 cm').

Only 21% of students reported that the image in the less accessible version of the question increased the ease of understanding. More than half of students (58%) preferred the version without the image. Some students suggested that the image was not informative and some of those who attempted this question version reported that they did not use the image.

Another interesting comment that arose was that highlighting important words with bold font style in the more accessible version of

Table 3: Frequencies of responses regarding Question 2 (Characteristics)

Was the question easy to understand?	V1 Less accessible	V2 More accessible		
Yes	17 (61%)	21 (72%)		
No	7 (25%)	2 (7%)		
Unclear/mixed	4 (14%)	6 (21%)		
Context of two sisters (with/without) – which is easier to understand?	**V1 – LA** (context)	**V2 – MA** (no context)	**No difference**	**Unclear/ mixed**
Frequency	6 (11%)	33 (58%)	7 (12%)	11 (19%)
Image (with/without) – which is easier to understand?	**V1 – LA** (image)	**V2 – MA** (no image)	**No difference**	**Unclear/ mixed**
Frequency	12 (21%)	33 (58%)	2 (4%)	10 (18%)

the question was useful. This is relevant to accessibility and part of OCR's usual formatting style (but is not one of the themes that the research set out to investigate).

Question 3

Question 3 was based around a graph of how world energy use (or demand) has changed over time. The graph showed different energy types and asked students how much the total world's energy use (or demand) had increased between certain years. There were differences in the wording and the graph between the question versions. The perceived understandability of this question was relatively low, with only about half of the students reporting that the question was easy to understand, regardless of the version they attempted.

Table 4: Frequencies of responses regarding Question 3 (Energy graph)

Was the question easy to understand?	V1 More accessible	V2 Less accessible		
Yes	14 (50%)	14 (48%)		
No	8 (29%)	9 (31%)		
Unclear/mixed	6 (21%)	6 (21%)		
Language (clarity of information) – which is easier to understand?	**V1 – MA** (extra sentence before graph, includes 'approximately')	**V2 – LA** (without extra sentence, excludes 'approximately')	**No difference**	**Unclear/ mixed**
Frequency	24 (42%)	8 (14%)	13 (23%)	12 (21%)
Vocabulary (use/demand) – which is easier to understand?	**V1 – MA** ('energy use')	**V2 – LA** ('energy demand')	**No difference**	**Unclear/ mixed**
Frequency	26 (46%)	1 (2%)	30 (53%)	0
Graph – which is easier to understand?	**V1 – MA** (larger graph with fewer energy types)	**V2 – LA** (smaller graph with more energy types)	**No difference**	**Unclear/ mixed**
Frequency	41 (72%)	2 (4%)	9 (16%)	5 (9%)

The two versions of the question differed in terms of the introductory text provided before the graph (the more accessible version contained an extra sentence intended to provide greater clarity about the categories in the graph) and in the way that the students were asked to provide the amount of energy use increase (the more accessible version included the word 'approximately'). In terms of these features, the more accessible version was considered easier to understand by 42% of interviewees (compared with 14% who thought the other version was easier to understand in this respect). Some students thought that 'approximately' indicated that their response did not need to be exact[3], though a smaller number of students reported that the word 'approximately' did not make a difference or that the question was simpler without it. In terms of other text differences, some students felt that the extra sentence before the graph (in the more accessible version) provided useful information, whilst others implied that having fewer words was an advantage of the less accessible version.

The question used the phrase 'energy use' or 'energy demand'. The phrase 'energy use' (more accessible version) was seen as easier to understand than 'energy demand' by 46% of interviewees. Only one student preferred the phrase 'energy demand'. That said, many students (53%) reported that it made no difference whether the word 'use' or 'demand' was used.

The majority of students (72%) found the larger graph showing fewer energy types (more accessible version) easier to understand and use. Students commented that the bigger graph was clearer and that showing fewer energy types made the graph less confusing.

Question 4

Question 4 was about a food chain involving oilseed rape. Students were asked to complete a pyramid of biomass and then to calculate the efficiency of biomass transfer from the oilseed rape to honeybees. Question 4 was included to evaluate the influence of the amount of detail provided. The less accessible version contained additional contextual detail (about human use of the oil). Both versions of the question were easy to understand according to most students (over 60% for both versions).

When asked to compare the question versions in terms of context, the majority of students (74%) preferred the shorter context (more accessible version). Students typically justified their choice by saying

Table 5: Frequencies of responses regarding Question 4 (Food chain)

Was the question easy to understand?	V1 Less accessible	V2 More accessible		
Yes	18 (64%)	20 (69%)		
No	7 (25%)	2 (7%)		
Unclear/mixed	3 (11%)	7 (24%)		
Context – which is easier to understand?	**V1 – LA** (detailed context)	**V2 – MA** (shorter context)	**No difference**	**Unclear/ mixed**
Frequency	3 (5%)	42 (74%)	9 (16%)	3 (5%)

3. The mark scheme rewarded answers that were correct to the nearest whole number so presumably the word 'approximately' was intended to indicate that responses did not need to be highly accurate.

that the additional information in the less accessible version was irrelevant to answering the question and that having less information to read is usually beneficial, especially under the time-constrained conditions of an exam.

Similarly to Question 2, several students commented that the highlighting of key words using bold font style (more accessible version) was useful.

Question 5

Question 5 was set in the context of a student watching a ball game and seeing the ball being hit before hearing the sound. Candidates were asked to describe the measurements the student would need to find the speed of sound. The less accessible version included a drawing of the student watching the game, whilst the more accessible version did not include an image. Question 5 was used to explore the influence of a non-essential visual resource on accessibility. More than half of the students felt that the version of the question that they attempted was easy to understand.

Table 6: Frequencies of responses regarding Question 5 (Ball game)

Was the question easy to understand?	**V1 More accessible**	**V2 Less accessible**		
Yes	15 (54%)	19 (66%)		
No	8 (29%)	5 (17%)		
Unclear/mixed	5 (18%)	5 (17%)		
Image (with/without) – which is easier to understand?	**V1 – MA** (no image)	**V2 – LA** (image)	**No difference**	**Unclear/ mixed**
Frequency	17 (30%)	29 (51%)	9 (16%)	2 (4%)

In contrast to the findings for Question 2, about half of the students (51%) expressed a preference for having the image of the ball game (in the less accessible version) rather than having no image (more accessible version). This was most commonly justified by the students in terms of the image helping to visualise the context of the question. However, nearly a third of students (30%) preferred the version of the question without the image, often suggesting that the image was not useful and that all the information was provided in the text.

Question 6

Question 6 was about a student conducting a titration experiment with an acid and an alkali (see Figure 1). Candidates were asked to describe and explain how the student could improve the experiment to get a more accurate result. Question 6 contained multiple accessibility-related differences between the two versions of the question, including differences in wording, presentation of contextual information (bullet points) and the provision of an additional image.

Most students who sat the more accessible version of the question (66%), found the question easy to understand. In contrast, less than half (46%) of students who sat the less accessible version reported that the question was easy to understand.

Of the 57 interviewed students, 56% found the language used in the more accessible version of this question easier to understand than that in the less accessible version. Note that some students confused wording and layout differences (i.e., bullet points), hence the relatively large proportion of students (37%) classified as 'unclear/mixed' for these features of Question 6.

The more accessible version of Question 6 used bullet points to explain the experiment. Most students (72%) reported that this version of the question was easier to understand than the alternative version, which did not use bullet points. Students commented that the less accessible version was more confusing, whereas bullet points presented the information clearly and were easier to follow.

The less accessible version of the question included a three-part diagram, which was reduced to two parts in the more accessible version (see Figure 1). Contrary to expectations, 44% of students thought that the three-part diagram was easier to understand whereas only 25% of students preferred the two-part diagram. Some students explained that the three-part diagram logically shows the steps of the experiment whilst the diagram in the other version missed out the first step.

Table 7: Frequencies of responses regarding Question 6 (Titration)

Was the question easy to understand?	**V1 Less accessible**	**V2 More accessible**		
Yes	13 (46%)	19 (66%)		
No	12 (43%)	7 (24%)		
Unclear/mixed	3 (11%)	3 (10%)		
Language (clarity of information) – which is easier to understand?	**V1 – LA** (later steps in method)	**V2 – MA** (main steps in method)	**No difference**	**Unclear/ mixed**
Frequency	0	32 (56%)	4 (7%)	21 (37%)
Layout – which is easier to understand?	**V1 – LA** (without bullet points)	**V2 – MA** (with bullet points)	**No difference**	**Unclear/ mixed**
Frequency	0	41 (72%)	0	16 (28%)
Diagram – which is easier to understand?	**V1 – LA** (three-part diagram)	**V2 – MA** (two-part diagram)	**No difference**	**Unclear/ mixed**
Frequency	25 (44%)	14 (25%)	14 (25%)	4 (7%)

Question 7

Question 7 was about the forces acting on a trolley on a ramp. The scenario was explained (partly by a diagram) and students were asked to calculate the gravitational potential energy transferred (part a) and then to give a best estimate of the distance travelled based on five readings (part b). Question 7 was selected to evaluate the importance of grammatical structure, the order of answer options (numerical) and unit presentation. This question appeared to be understood by the majority of students, with 79% of students who sat the more accessible version of the question and 62% of students who sat the less accessible version claiming that they found the question easy to understand.

When asked to compare the versions of the question, the majority of students (75%) reported finding the simpler sentence structure in the more accessible version of the question easier to understand than the longer sentence in the other version. Students often justified their

choice by saying that the lengthy sentence could be confusing and separating out the value to be used for gravitational field strength (by splitting the sentence into two) meant that the information was clearer.

Part (b) of Question 7 was a multiple choice question where students answered by ticking a box. A simpler instruction regarding ticking the box was used in the more accessible version. Around half of the interviewed students (49%) felt that this difference in the wording made no difference to ease of understanding. Students typically commented that the meaning of the instructions was the same. However, more students preferred the shorter instruction (33%) than the number who preferred the longer instruction (14%).

The order of the answer options for part (b) was numerical in the more accessible version of the question and random in the less accessible version. Whilst half of the students (51%) suggested that the order of the answer options did not affect the ease of understanding the question, almost all of the remaining students (47%) expressed a preference for numerical order.

The final feature that was explored using this question was the presentation of the abbreviation for metres in a table. The 'm' for metres was presented in brackets in the more accessible version of the question and after a slash symbol in the less accessible version. Over 60% of students felt that the units were easier to understand when presented in brackets. Some students commented that they were more familiar with brackets being used to display units or that the slash could be misinterpreted (e.g., as a symbol for 'divide').

Table 8: Frequencies of responses regarding Question 7 (Trolley on a slope)

Was the question easy to understand?	**V1 More accessible**	**V2 Less accessible**		
Yes	22 (79%)	18 (62%)		
No	5 (18%)	7 (24%)		
Unclear/mixed	1 (4%)	4 (14%)		
Language (grammatical structure: general) – which is easier to understand?	**V1 – MA** (shorter instruction for part (a), other simpler sentences)	**V2 – LA** (longer instruction for part (a), other more complex sentences)	**No difference**	**Unclear/ mixed**
Frequency	43 (75%)	1 (2%)	7 (12%)	6 (11%)
Language (grammatical structure: tick instruction) – which is easier to understand?	**V1 – MA** ('Tick **one** box')	**V2 – LA** ('Put a tick in the **one** correct box.')	**No difference**	**Unclear/ mixed**
Frequency	19 (33%)	8 (14%)	28 (49%)	2 (4%)
Order – which is easier to understand?	**V1 – MA** (number order)	**V2 – LA** (random order)	**No difference**	**Unclear/ mixed**
Frequency	27 (47%)	1 (2%)	29 (51%)	0
Units – which is easier to understand?	**V1 – MA** ('(m)')	**V2 – LA** ('/m')	**No difference**	**Unclear/ mixed**
Frequency	36 (63%)	0	17 (30%)	4 (7%)

Question 8

Question 8 described a student investigating the effect of acid rain on seed growth by observing how many seeds germinate in the presence of solutions of different pH. Candidates were asked to give a factor that should be kept the same during the investigation and to describe what the results indicate. Question 8 was included to evaluate the influences of using bullet points to present contextual information and of the alignment of figures and tables (left-aligned versus centred). Around 60% of students attempting each version of the question reported that the question was easy to understand.

There was an overwhelming preference for bullet point presentation of the context, with 74% of students claiming that the more accessible version (with bullet points) was easier to understand. Students often commented that the bullet points looked clearer and identified the key information needed for answering the question.

Most students (70%) felt that the alignment of the figure and table did not affect how easy the question was to understand. For those students who expressed a preference, the version with the left-aligned figure and table was chosen marginally more often (18%) than the version with the figure and table positioned centrally (12%).

Table 9: Frequencies of responses regarding Question 8 (Acid rain/seed germination)

Was the question easy to understand?	**V1 Less accessible**	**V2 More accessible**		
Yes	18 (64%)	17 (59%)		
No	4 (14%)	2 (7%)		
Unclear/mixed	5 (18%)	8 (28%)		
N/A – did not reach this question/ran out of time	1 (4%)	2 (7%)		
Layout – which is easier to understand?	**V1 – LA** (without bullet points)	**V2 – MA** (with bullet points)	**No difference**	**Unclear/ mixed**
Frequency	1 (2%)	42 (74%)	10 (18%)	4 (7%)
Alignment of figure and table – which is easier to understand?	**V1 – LA** (centre-aligned)	**V2 – MA** (left-aligned)	**No difference**	**Unclear/ mixed**
Frequency	7 (12%)	10 (18%)	40 (70%)	0

Summarised findings for each accessibility theme

Table 10 summarises the findings for each accessibility theme explored. Findings that were counter to expectations are shown in red. Neutral findings (where most students felt the feature made no difference to the ease of understanding and where there was no general direction of preference amongst those who did express a preference) are shown in blue.

Discussion

The aim of this research was to investigate students' perceptions of exam questions with and without OCR's accessibility principles applied. For most of the question features that were explored in this study, student perceptions of accessibility tended to align with expected effects on

Table 10: Summarised findings by accessibility theme

OCR principle (OCR, 2018a)	*Theme explored*	*Summary of findings (red text indicates findings that were counter to expectation, blue text indicates findings where views tended to be neutral)*
2	Language	• When given the choice between a simpler term ('use') and slightly more complex vocabulary term ('demand'), almost all students either found the simpler term easier to understand (46%) or felt the term made no difference (53%) (Q3); • Students tended to find question versions with simpler sentence structures easier to understand, though the strength of this finding varied (Q7 general, Q7b); • Text changes intended to aid clarity (but which did not involve a difference in grammatical complexity) were reported by more students to be easier to understand. (These versions of questions sometimes had a higher word count) (Q3, Q6).
4	Presentation of context	• Students tended to consider questions with shorter contexts or no context easier to understand (Q2, Q4); • Nearly three-quarters of students found question versions that used bullet points to set out the steps in a process or method easier to understand than question versions that did not (Q6, Q8).
8	Order of MCQ answer options	• For MCQ answer options involving phrases, most students felt the order made no difference (Q1); • For numerical MCQ answer options, just over half of students felt that the order made no difference and a little under half of the students felt that numerical order was easier to understand (Q7b).
10	Units presented in brackets for tables	• Most students felt that showing units in brackets was easier to understand than the units being preceded by a slash symbol. Others felt it made little difference, but none preferred the slash symbol (Q7b).
13	Visual resources	• Non-essential images: o For one question with a non-essential image, over 50% of students felt that the question was easier to understand without the image whilst around 20% preferred having the image (Q2); o For another question with a non-essential image, around half of students reported that the question was easier to understand with the image whilst around 30% preferred the version without the image (Q5); o For a question where an extra part to the diagram showed a preceding step in an experiment, 44% of students preferred the three-part diagram whilst 25% preferred the two-part diagram (Q6). • Over 70% of students felt that a larger graph showing fewer different substances was easier to understand (Q3).
14	Left alignment	• Most students (70%) felt that the alignment of a figure and table (left or centred) made no difference to understanding the question. A few students expressed a preference for one or the other (Q8).

accessibility but there were some exceptions. We reflect below on the findings for each accessibility theme.

Language

Differences in the language used, such as vocabulary and grammatical structure, affected perceived accessibility in the expected direction. However, for the vocabulary issue and one of the grammatical complexity issues explored there were fairly high numbers of students who felt that the language differences did not affect the ease of understanding. This may suggest that these changes were helpful to those students with slightly weaker language skills but were less necessary for others. In the case of vocabulary, the influence of changes will depend on the specific words used and how familiar the words are to the general student population and to individuals within that population. Where changes did not appear to help all students but did reportedly help a proportion of students (and did not seem to hinder others), there is still a strong argument for implementing such changes in order to reduce risks that language skills negatively affect performance for some students (where it is not the intention to assess language skills).

Presentation of context

The findings relating to context were in line with expected effects. Using bullet points to set out steps in a method or process appeared to be helpful to most students in understanding contextualised questions. This is interesting given that past research has produced mixed findings on the effect of bullet points on accessibility (Crisp, Johnson, & Novaković, 2012; Kettler et al., 2012). Reducing unnecessary detail in a context (Q4) and removing a context in a question where the context potentially caused confusion (Q2) tended to help students to understand the question, according to the interviewees. However, it should be noted that good contexts can usefully facilitate the assessment of certain kinds of skills (Ahmed & Pollitt, 2007) and the current findings should not be interpreted to mean that removing or minimising context is always going to enhance accessibility or is always the appropriate choice in terms of assessing the skills of interest. Nonetheless, it appears that it may be advisable to avoid including unnecessary contextual information.

Order of answer options in multiple choice questions

If anything, students tended to report that positioning response options for multiple choice questions in numerical order was easier to understand than having options presented in random order. That said, over half of the students felt that the order made no difference.
As mentioned earlier, where a change appears to aid accessibility for more students than it hinders, this change is probably good practice even if it makes little difference to some students. The majority of interviewees felt that presenting response options in alphabetical order did not make a difference to the ease of understanding Question 1.
This may have been partly a result of the response options being short sentences and there being no relationship between the meaning of these sentences and the order of their presentation (either alphabetical or random). Other multiple choice questions could have such a relationship and, thus, alphabetical order might benefit students. In any case, the current research did not suggest that alphabetical order was a hindrance to students and potentially still serves OCR's intended purpose of using alphabetical and numerical order to avoid the order of the options

potentially giving away the correct answer. Additionally, using alphabetical or numerical order is logical and tends to be considered good practice (e.g., Moncada & Moncada, 2010).

Units presented in brackets for tables

In line with OCR's expectations about the effect of question features, presenting the abbreviation for metres in brackets was felt by most students to be easier to understand, suggesting that this does aid accessibility. This style was reportedly more familiar and less likely to cause confusion than using a slash symbol.

Visual resources

OCR's principles set out that images and diagrams (and data) will "only be used where they genuinely support what is required in the question" to avoid "distracting images for the students that do not help them understand what is required" (OCR, 2018a, p.7). This is a sensible decision given that visual resources in questions are salient, can dominate students' thinking and, thus, can be misleading if the information they contain is not genuinely relevant (Crisp & Sweiry, 2006). Additionally, Kettler et al. (2012) argued that introducing non-essential images is likely to increase cognitive load and divert students' attentional resources from the focus of the question.

For two questions in the current research, non-essential images were removed in the more accessible version. Findings for one question (Q2) were in line with expectations, with more students (58%) reporting that the version without the image was easier to understand (though it should be noted that 20% preferred the illustrated version). For the other question with a non-essential image (Q5), the opposite pattern was found with more students finding the less accessible version with the image easier to understand (51%) (though 30% preferred the unillustrated version). The findings were also counter to expectations for a further question (Q6); more students preferred a three-part diagram (preferred by 44%) to a two-part diagram (preferred by 25%) where an initial step in an experiment was not shown. These rather mixed findings suggest that the exact nature of the image and its relation to the question could be affecting views on accessibility. One hypothesis would be that images appearing to be more diagrammatic or more informative about the scenario are more likely to improve understanding of the question. This would be consistent with the cartoon-like image in Question 2, which gave no additional information, being least appreciated. This aligns with findings from Crisp and Sweiry (2006) suggesting that students have appropriate expectations regarding which aspects of a visual resource are important and relevant. OCR's principle to exclude visuals that do not support answering the question is still sound, but the current findings emphasise that decisions around the inclusion of visual resources should be made on a case-by-case basis taking into account the nature of the specific visual and how it might potentially support interpretation of the question. This is consistent with OCR's current practice.

With regard to the clarity of visuals, the findings support the notion that it is important to ensure that any visual resources are clear and easy to interpret, given that the larger graph showing fewer substances in the more accessible version of Question 3 was reportedly easier to understand, according to most of the interviewed students.

Left alignment

To be consistent with the principles applied for modified papers, OCR's accessibility principles set out that visual resources will be left aligned (unless students are required to work with the resource in a way that makes having space around the resource helpful). Left alignment is thought to be easier to understand for those with dyslexia or certain visual impairments (Evett & Brown, 2005). For the group of students interviewed in the current research, most students felt that the alignment of the figure and table in Question 8 did not affect how easy the question was to understand. Amongst those students who expressed a preference, there was no general trend in the direction of their views. Whilst the principle to left align visual resources did not appear to aid the sample of students interviewed, it also did not hinder them so it would still seem appropriate to apply this accessibility principle on the grounds that it may help those with visual impairments and dyslexia.

Limitations

The current research has some potential limitations. During interviews, students were encouraged to discuss each question feature relating to accessibility in turn and in most cases separate comments on different accessibility principles were gathered. Nonetheless, it was evident that different features of the questions sometimes interacted with one another and the impact of individual principles could not always be assessed. Each accessibility theme was explored in relation to a small number of questions and it is possible that findings might have been different for a similar feature appearing in a different question, depending on other features of the question. In addition, as the students were interviewed in pairs, their opinions could have been influenced by their peers. However, as the assignment of test versions to students was random, it is unlikely that this would have led to a systematic bias in responses.

Conclusion

When addressing the notion of accessibility, the focus is on the target user's experience and giving them a fair opportunity to attempt the questions presented in order to show their ability in the construct(s) of interest. An additional aim of this is to provide a more positive experience for the students in terms of being able to engage with the questions. However, there is a distinction between perceived accessibility and the actual effect on performance, which should be kept in mind when interpreting the findings from the current research.

For most of the accessibility themes explored, student perceptions of the ease of understanding different versions of questions were in line with expectations about effects on accessibility. For two accessibility themes, the findings were neutral. For one accessibility theme, the removal of a non-essential visual resource (or part of one), there were varying effects on perceived accessibility. Whilst the effects for visuals were mixed, other evidence (Crisp & Sweiry, 2006; Kettler et al., 2012) supports the notion that visuals which do not provide useful information are best avoided, and it would seem reasonable to retain this accessibility principle.
In conclusion, the students' views gathered in this research suggest that the accessibility principles that we investigated are appropriate and should continue to be applied to help ensure students can understand and access future exam questions.

Acknowledgement

We would like to thank the teachers and students who helped us with this research for their time and enthusiasm.

References

Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: An experimental investigation of focus. *Assessment in Education: Principles, Policy and Practice, 14*(2), 201–232.

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy and Practice, 18*(3), 259–78.

Baker, W.H. (2001). HATS: A design procedure for routine business documents. *Business Communication Quarterly, 64*(2), 66–78.

Beddow, P.A., Elliott, S.N., & Kettler, R.J. (2013). Test accessibility: Item reviews and lessons learned from four state assessments. *Education Research International, 2013*, Article ID 952704, 1–12.

Beddow, P.A., Kurz, A., & Frey, J.R. (2011). Accessibility theory: Guiding the science and practice of test item design with the test-taker in mind. In S.N. Elliott, R.J. Kettler, P.A. Beddow, & A. Kurz (Eds), *Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice and policy* (pp.163–182). New York: Springer.

Chelesnik, A.L. (2009). *The impact of test typography on student achievement, anxiety, and typographic preferences*. Unpublished Masters Thesis. California State University, San Marcos. Available from: https://csusm-dspace.calstate.edu/handle/10211.3/114688 (retrieved 4th November 2019).

Crisp, V. (2011). Exploring features that affect the difficulty and functioning of science exam questions for those with reading difficulties. *Irish Educational Studies, 30*(3), 323–343.

Crisp, V., Johnson, M., & Novaković, N. (2012). The effects of features of examination questions on the performance of students with dyslexia. *British Educational Research Journal, 38*(5), 813–839.

Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research, 48*(2), 139–154.

Evett, L., & Brown, D. (2005). Text formats and web design for visually impaired and dyslexic readers – Clear Text for All. *Interacting with Computers, 17*(4), 453–472.

Ketterlin-Geller, L.R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice, 27*(3), 3–16.

Kettler, R.J., Dickenson, T.S., Bennett, H.L., Morgan, G.B., Gilmore, J.A. et al. (2012). Enhancing the accessibility of high school science tests: A multistate experiment. *Exceptional Children, 79*(1), 91–106.

Lonsdale, M. dos S., Dyson, M. C., & Reynolds, L. (2006). Reading in examination-type situations: The effect of text layout on performance. *Journal of Research in Reading, 29*(4), 433–453.

Moncada, S.M., & Moncada, T.P. (2010). Assessing student learning with conventional multiple-choice exams: Design and implementation considerations for business faculty. *International Journal of Education Research, 5*(2), 15–29.

OCR. (2018a). GCSE (9–1) *Gateway Science: Exploring our question papers*. Cambridge: OCR. Retrieved from: https://www.ocr.org.uk/Images/462559-exploring-our-question-papers-gateway-science.pdf (retrieved 4th November 2019).

OCR. (2018b). GCSE (9–1) *Twenty First Century Science: Exploring our question papers*. Cambridge: OCR. Retrieved from: https://www.ocr.org.uk/Images/462607-exploring-our-question-papers-twenty-first-century-science.pdf (retrieved 4th November 2019).

OECD. (2009). *Learning mathematics for life: A perspective from PISA (Programme for International Student Assessment)*. Paris: OECD Publishing.

Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands of examination syllabuses and question papers. In P. Newton, J-A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds), *Techniques for monitoring the comparability of examination standards* (pp.166–206). London: Qualifications and Curriculum Authority.

QCA. (2004). *The statutory regulation of external qualifications*. London: Qualifications and Curriculum Authority.

QCA. (2005). *Fair access by design: Guidance for awarding bodies and regulatory authorities on designing inclusive GCSE and GCE qualifications*. London: Qualifications and Curriculum Authority.

Spalding, V. (2009). *Is an examination paper greater than the sum of its parts? A literature review of question paper structure and presentation*. Manchester: AQA Centre for Education Research and Policy.

Using corpus linguistics tools to identify instances of low linguistic accessibility in tests

David Beauchamp and **Filio Constantinou** Research Division

Introduction

Assessment is a useful process as it provides teachers and other stakeholders (e.g., parents, government, employers) with information about students' competence in a particular subject area. However, for the information generated by assessment to be useful, it needs to support valid inferences. One factor that can undermine the validity of inferences from assessment outcomes is the language of the assessment material. For instance, if a Mathematics test question contains complex vocabulary and/or grammar, it might prevent students from demonstrating their true mathematical knowledge and skills. This may result in teachers and other stakeholders drawing inaccurate inferences from the test scores. Students who are not native speakers of the target language are more likely to be disadvantaged by assessment material that displays low levels of linguistic accessibility. In an attempt to support teachers and test developers in designing linguistically accessible assessment material, this study explored practical ways of investigating the complexity of test questions

both at the level of vocabulary (lexical complexity) and grammar (syntactic complexity).

The starting point of this research was the shortcomings of traditional measures of linguistic accessibility, or readability, and their limited applicability to test questions. For example, traditional readability measures often assume that longer words are more difficult to comprehend (see Lenzner, 2014). However, in the context of assessment, such words are normally subject-specific technical terms (e.g., *microorganism, photosynthesis*) with which students are expected to be familiar, as they are part of the construct that is being assessed. Also, traditional readability measures tend to be based upon continuous prose and fully formed sentences and, as a result, are not well-suited for measuring the readability of texts that do not fit this format, especially multiple-choice questions for example. Furthermore, readability measures that are based on sentence length and text length do not consider the different cognitive challenges that various syntactic structures pose on readers (Lenzner, 2014).

In response to these shortcomings, alternative ways of investigating the linguistic accessibility of assessment materials were explored. These involved undertaking lexical and syntactic analyses of test questions in an automated manner using software packages typically employed in the field of corpus linguistics (for a definition of corpus linguistics, see the Method section below). To our knowledge, this study represents one of the first attempts to identify instances of low linguistic accessibility in assessment material using corpus linguistics methods. In this study, accessibility is understood as "the degree to which a test and its constituent item set permit the test taker to demonstrate his or her knowledge of the target construct [and] is conceptualized as the sum of interactions between features of the test and individual test taker characteristics" (Beddow, Elliott, & Kettler, 2013, p.1).

Lexical complexity

The issue of lexical complexity, or lexical sophistication, in testing is often discussed in the context of the assessment needs of second language speakers. Second language speakers constitute a particularly vulnerable group as they are assessed via a language that is different from their mother tongue. In the context of high-stakes testing, characteristic is the study by Shaw and Imam (2013) that sought to identify, among other linguistic resources, the vocabulary needed by non-native English speakers to complete IGCSEs in History, Biology and Geography successfully. The lexical resources were then classed according to the Common European Framework of Reference for Languages (CEFR), an international scale that describes language competence (Council of Europe, 2018).

An important distinction to make when considering the challenge that vocabulary poses to a test taker is that between *content-obligatory language* and *content-compatible language* (Cloud, Genesee, & Hamayan, 2000). The former includes technical, subject-specific language needed to understand and respond to test items (e.g., *photosynthesis* and *Reformation* for Biology and History respectively), while the latter is a foundation of more common, non-subject-specific language (e.g., *plants* and *social development* for Biology and History respectively). This distinction is important because when identifying instances of lexical complexity which may compromise accessibility, one must discount what is likely to be content-obligatory vocabulary, the learning and use of which makes up part of the construct to be assessed.

Research concerned with the lexical complexity of texts has involved the compilation of vocabulary level lists that have been used in lexical analysis software, such as the RANGE program and AntWordprofiler, and in tests designed to assess learners' lexical knowledge such as the Vocabulary Size Test (see Anthony, 2013; Bauer & Nation, 1993; Beglar & Nation, 2007; Nation, 2018; Webb & Nation, 2008). The most extensive vocabulary level lists are based upon language use in the British National Corpus (BNC) and Corpus of Contemporary American English (COCA) (see Nation, 2018). Each level of these lists consists of vocabulary derived from 1000 word families (a word family is vocabulary based around a root word such as *give*, and its derivatives such as *gives*, *giving*, *given*). In particular, Level 1 consists of vocabulary based upon the first 1000 word families an English language learner is likely to encounter, Level 2 is based upon the next thousand word families, and so on. The vocabulary grows progressively more obscure through 29 levels. Table 1 below provides examples of words across the different levels, as found in the BNC/COCA vocabulary lists and in A Level Biology examination papers.

Table 1: Examples of words across levels, as found in the BNC/COCA vocabulary lists and in A Level Biology examination papers

Vocabulary list level	*Examples from the BNC/COCA vocabulary lists*	*Examples from A Level Biology examination papers*
1	offer (offers, offered); stay (stays, stayed, staying); carry (carries, carried, carrier)	what; show; main; that; student
2	access (accesses, accessed, accessible); fry (fries, fried, fryer)	section; indicator; repeated
3	abandon (abandons, abandoned, abandoning); collapse (collapses, collapsed, collapsing); promote (promotes, promoted, promoters)	vessel; theory; evolved
4	abnormal (abnormality, abnormalities, abnormally); prestige (prestiges, prestigious); subsidiary (subsidiaries, subsidiarity)	graph; acid; interval
5	accessory (accessorise, accessorised, accessories); burgle (burgled, burglar, burglaries); lurk (lurks, lurked, lurking)	saturate; niche; botany
6	abduct (abducted, abducting, abduction); clutter (clutters, cluttered, cluttering); incubate (incubates, incubated, incubation)	chromosome; receptor; aquatic
7	abate (abated, abatement, abating); ludicrous (ludicrously, ludicrousness); throng (throngs, thronged, thronging)	tentacle; amphibian; viral
8	abstinence (abstinences); orator (oratories, orators, oratory); paraphrase(paraphrases, paraphrased, paraphrasing)	catalyse; yoga; biodiversity
9	abyss (abysses, abyssal); denominator (denominators)	photosynthesis; collagen; microorganism
10+	adage (adages); libertine (libertines); portcullis	habituate; hydrolysis; glycaemic

Syntactic complexity

Syntactic complexity is concerned with linguistic structures above the level of the individual word (e.g., clauses, sentences). Syntactically complex texts can increase cognitive load and thus undermine accessibility by placing the barrier of good reading skills and good working memory before the construct to be tested.

Research into linguistic accessibility has identified syntactic features that can affect comprehension. Štajner, Evans, Orasan and Mitkov (2012) reported that subordinating phrases, coordinating phrases, infinitives and prepositional phrases as grammatical structures were generally associated with a lower degree of readability on the Flesch scale (see Table 2 for examples of some of these structures). In a similar vein, Ariel (2001) has developed a spectrum of linguistic accessibility markers which ranges from *low accessibility* markers (e.g., long descriptions, long noun phrases) to *high accessibility* markers (e.g., pronouns, noun omission), with less linguistic material generally being more favourable for cognitive processing. The level of conceptual content within sentences has also been considered as a factor that may affect readability. For instance, Feng, Jansche, Huenerfauth and Ehladad (2010) found that the number of general nouns and named entities in a text, also known as entity-density, performed well as a readability measure, with greater *entity-density* indicating lower readability.

Subordinating phrases in the form of nested clauses (clauses embedded within other clauses) are considered to increase linguistic complexity, as they require greater mental effort on the part of the reader to be successfully processed (Gibson, 1998; Miller & Isard, 1964). However, it is not only the presence of certain syntactic features that can affect the complexity of a sentence. The position of such features within the sentence can also have implications for complexity and, by extension, accessibility. In a study on survey question difficulty, Lenzner (2014) points to the difficulty in processing left-branching structures. These are structures which contain considerable linguistic material in the form of clauses, phrases or other modifiers before the main verb is reached. The need to process this linguistic material prior to encountering the main verb in the sentence tends to increase the demand on working memory. The sentences below exemplify the difference between left-branching (a) and right-branching (b) structures:

> *(a) How likely is it that if a law was considered by parliament that you believed to be unjust or harmful, you, acting alone or together with others, would try to do something against it?*
>
> *(b) How likely is it that you, acting alone or together with others, would try to do something against a law that was considered by parliament and that you believed to be unjust or harmful?*
>
> Lenzner (2014, p.685)

Concerning examination papers, much work has focused on the presence of linguistic complexity in Mathematics papers, probably owing to the risk that excessive language processing poses to an assessment in which the target construct is essentially a non-linguistic one. A range of studies have been carried out examining the effects of aspects of syntactic complexity on the performance of EAL (English as an Additional Language) students in Mathematics tests, using candidate interviews, DIF (Differential Item Functioning) statistics and regression analyses (e.g., Martiniello, 2008; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006; Wolf et al., 2008). These studies focused on various markers of syntactic complexity in Mathematics papers including sentence length, item length, noun phrase length, and the presence of prepositional phrases, participles and multiple and relative clauses. Their results showed that the effects of syntactic complexity on candidate performance were limited or inconclusive. Table 2 below illustrates how these syntactic features are manifested in A Level Biology examination papers.

Table 2: Features contributing to syntactic complexity as manifested in A Level Biology examination papers

Syntactic feature	*Example*
Subordinating clause	Complete Table 1 by putting a tick in a box **if the structure is present in the type of cell**. The reserve managers chose a high temperature **because this causes the young lizards to hatch more quickly**. **Although a moss plant has no vascular tissue,** water still moves through the plant from the root-like structures to the leaves.
Passive structure	The volunteers **were asked** to record three symptoms.
Prepositional adjunct	The circles in Figure 1 represent the hierarchy **of taxonomic groups for the classification** shown **in Table 1**.
"to" infinitive	He used a pH meter **to record** pH.
Past participle phrase	The table below shows the vitamin C content of sauerkraut and cabbage, **treated in different ways**.
Present participle phrase	**Using a genetic diagram**, find the probability that the next child born to parents 3 and 4 would be affected by moyamoya.
Relative clause	The photograph below shows packaging pellets made from thermoplastic starch, **which is produced from corn starch**. This investigation was carried out in a university laboratory, using species of bacteria **that cause disease in humans.**

With a view to making mathematics items more linguistically accessible to candidates, Abedi and Lord (2001) simplified verb phrases, conditional clauses, relative clauses, question phrases and abstract representations. They found that EAL and non-EAL students alike made small but statistically significant improvements on simplified items, as did students from low socio-economic backgrounds. Additionally, they found that items that had been simplified were more likely to be selected by candidates when a choice was given.

Method

To investigate linguistic accessibility in assessment material, three corpora of examination papers were compiled. The examination papers were obtained from three A Level subjects that represented different disciplines: Biology, Business Studies and History. The papers were developed by three major examination boards in England and were taken by students in the UK between 2015 and 2017. Each corpus was approximately 15,000 words long and comprised several hundred examination questions, covering a wide range of examples of examination questions typically encountered by candidates. The three corpora were explored using software packages commonly employed in corpus linguistic studies.

Corpus linguistics can be defined as a method of analysing "the actual patterns of use in natural texts" (Biber, Douglas, Conrad, & Reppen, 2004, p.4). It involves compiling large bodies of text, or corpora, and analysing them via specialist software to identify the presence, distribution and frequencies of various linguistic features. Analysing language use by means of corpus linguistics software, rather than manually, has certain advantages. These include (a) the capacity to analyse large amounts of text within a very short amount of time, and (b) the ability to identify trends that may be missed through an 'intuitive' reading by an individual. To our knowledge, to date, corpus linguistics software has not been used to investigate language use in assessment materials.

In this study, two corpus linguistics software packages were mainly used: AntWordProfiler (Anthony, 2013) and Multidimensional Analysis Tagger (Nini, 2015). The former was used for the lexical analysis, while the latter was used for the syntactic analysis.

AntWordProfiler: lexical analysis

AntWordProfiler is a software program which allows corpora of texts to be compared to imported word lists (Anthony, 2013). The software ranks the words in the texts according to their level of complexity (i.e., the inferred likelihood of a person knowing a word based upon the frequency of its use within a corpus of real language use). In this study, the BNC/COCA vocabulary level lists (see Nation, 2018) were used to provide a scale against which the vocabulary in the examination papers could be ranked. More specialised and technical vocabulary (e.g., scientific and historical terms) forms the content of the higher lists, while more commonplace, non-technical vocabulary forms the content of the lower lists. To frame these lists in a more widely known scale, Nation has provided an approximate classification of these vocabulary level lists based on the CEFR levels via personal communication (P. Nation, 21 September, 2018). This approximate classification is shown in Table 3 below.

As can be seen in Table 3, vocabulary which is present in lists 5 to 9 may not be known by candidates who are not "proficient" in English. As such, it could be viewed as representing a barrier to accessibility. On the other hand, vocabulary found in lists 10 and higher tends to be specialist or technical vocabulary that forms part of content-obligatory language and, as such, it is likely that it will have been encountered by candidates. However, it should be noted that this is not always the case. For instance, as can be seen in Table 1 above, there are examples of technical terms which are found in lists lower than level 10 (e.g., 'photosynthesis' which appears in list 9).

Multidimensional Analysis Tagger: syntactic analysis

Multidimensional Analysis Tagger (MAT) is a software package that analyses plain text files and uses a parts-of-speech (POS) tagger to identify and label syntactic features (Nini, 2015). The results of the analyses are then displayed in a table format. From these results, it is possible to isolate the presence and frequency of relevant syntactic features and structures across different texts. The syntactic features considered in this study are shown in Table 4 (see also Nini, 2015). They were chosen because: (a) they represent multiword structures which increase the linguistic material (and thus cognitive load) of the text; (b) they represent a variety of different semantic relations between entities; and (c) some of them have been shown in previous studies to affect text readability (see e.g., Štajner et al., 2012). The chosen features are not necessarily considered to be equal in the challenges they pose to readability.

It should be noted that MAT does not carry out the syntactic analysis at the level of the sentence or the clause, but only at the level of the provided text file. In this study, some of the analyses were carried out at the level of the subject corpus, while some others were carried out at the level of the item. Although the syntactic features considered in this study may have indicated the presence of syntactic complexity, the way in which the complexity was distributed among different sentences had to be identified through manual human analysis.

Table 3: Classification of BNC/COCA vocabulary level lists based on CEFR

CEFR level			*BNC/COCA vocabulary level lists*	
Proficient	C2	Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.	7000–9000 words	Lists 7–9+
	C1	Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.	5000–6000 words	Lists 5–6
Independent	B2	Has a good range of vocabulary for matters connected to his/her field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.	4000 words (2000–3000 high frequency words plus 1000–2000 relevant technical vocabulary)	List 4
	B1	Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel, and current events. Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics.	2000–3000 high frequency words	Lists 2–3
Basic	A2	Has a sufficient vocabulary for the expression of basic communicative needs. Has a sufficient vocabulary for coping with simple survival needs.	The most frequent 1000 word families	List 1
	A1	Has a basic vocabulary repertoire.	120 words and phrases from the survival vocabulary (=vocabulary needed for coping with simple survival needs)	List 1

Table 4: Syntactic features considered in this study (see Nini, 2015)

	Syntactic feature	*Example*
Additional clauses, beyond the simple structure of subject + verb + object	Causative adverbial clauses indicated by *because*.	The *business failed* ***because there was a lack of demand for the product***.
	Concessive adverbial clauses indicated by the words *although* and *though*.	***Although the snakes are venomous****, they rarely approach humans*.
	Conditional adverbial clauses indicated by the words *if* and *unless*.	*The campaign would be more successful* ***if it used targeted advertising***.
	Other adverbial subordinating clauses signalled by words such as *since*, *while* and *whereas*.	***Whereas the economy of the Northern states was increasingly industrial****, the economy of the Southern states remained predominantly agricultural.*
Passive structures		The journal ***is published*** biannually by the press.
Prepositional adjuncts		***In 1871****, Germany was unified* ***by Bismarck***.
"to" infinitives		*They agreed* ***to stop*** *selling the product after the lawsuit.*
Participles: past and present		***Built in a single week****, the house would stand for fifty years.* ***Stuffing his mouth with cookies****, Joe ran out of the door.*
Relative clauses	*Pied-piping* relative clauses: any preposition followed by *who*, *whom*, *whose*, or *which*.	*The manner* ***in which he was told***.
	That relative clauses in an object position.	*The dog* ***that I saw***.
	That relative clauses in a subject position.	*The dog* ***that bit me***.
	Sentence relatives: indicated by a punctuation mark followed by *which*.	*Bob likes fried mangoes,* ***which is disgusting***.
	What clauses	*I believed* ***what he told me***.
	Who relative clauses in an object position.	*The man* ***who Sally likes***.
	Who relative clauses in a subject position.	*The man* ***who likes popcorn***.

Findings

Key observations from the lexical and syntactic analyses are presented below.

Lexical analysis

There was variation in the level of lexical complexity that was observed across the three corpora. While the vast majority of vocabulary was indicated to be at an accessible level (89.7%–93.6% of vocabulary lay within levels 1 to 4), each subject corpus included examples of vocabulary of increasing complexity and obscurity which could potentially disadvantage some candidates, especially EAL ones. The Biology corpus displayed the highest proportion of language at levels 5 to 9 (4.3% as opposed to 1.3% in Business Studies and 1.2% in History), while the History corpus displayed the highest proportion of vocabulary at levels 10 and above. As mentioned earlier, according to Nation, vocabulary at levels 5 to 9 tends to correspond to vocabulary expected of second language speakers who are at CEFR levels C1 and C2 (i.e., "proficient level"). On the other hand, vocabulary at level 10 and above often indicates subject-specific vocabulary, or content-obligatory language (including proper nouns and dates). Characteristic examples of words which may disadvantage EAL students (i.e., words which are not subject-specific and are at level 5 or above) can be found in the following History items:

> *Assess which religious issue most* ***hindered*** *the development of […] in the period from […].*
>
> *Study all the Sources. Use your own knowledge to assess how far the Sources support the interpretation that the difficulty in finding a solution to the problems of […] was the* ***reluctance*** *of the […] to co-operate with […].*

More examples of non-subject-specific vocabulary at levels 5 to 9 that occurred in the examination papers analysed can be found in Table 5.

Table 5: Non-subject-specific vocabulary at levels 5 to 9 used in examination papers in Biology, Business Studies and History

Level	*Biology*
Level 5	*miniature, voyage, expel*
Level 6	*stranded, streamline*
Level 7	*rupture, tar, deduce*
Level 8	*dissociate, frill*
Level 9	*sheath*
Level	*Business Studies*
Level 5	*incur, mattress, morale, pier, ruthless, hawk, trailer, flop, goose, grooming, ignite, orphan, underestimate, abolish, brochure*
Level 6	*souvenir, outweigh, stout, hygiene, drawback, wasp, glossy, mentor*
Level 7	*bingo, scaffold, titan, ware*
Level 8	*gourmet, posh, fang, aptitude*
Level 9	*fizz, kiln*
Level	*History*
Level 5	*hinder, voyage, reluctance*
Level 6	*influx*
Level 7	*blunder, gravely, misplace*
Level 8	*hermit*
Level 9	–

Where there was uncertainty as to whether certain words were subject-specific or not, AntConc (a free concordancing and text analytics package) (see Anthony, 2018) was used to identify occurrences of these words in the respective syllabi. However, it should be acknowledged that the distinction between 'subject-specific' and 'non-subject-specific' vocabulary is not clear-cut and that some non-subject-specific words that are relatively rare in everyday discourse may also be encountered in specific classroom teaching (e.g., 'tar' in the context of health risks of smoking).

Syntactic analysis

As the analyses carried out via MAT showed, the three subject corpora displayed considerable differences in terms of their use of grammatical features that tend to contribute to syntactic complexity. For example, "to" infinitives were comparatively over-represented in Business Studies, suggesting a focus on verbs and actions. Similarly, passives predominated in Biology, suggesting a tendency towards more formal language and the reporting of processes. Although these observations indicate little in terms of the accessibility of individual items, they suggest differences in item construction across subjects.

Pairs of items which were similar in some respects (e.g., were obtained from the same subject; were of similar length) but had a comparatively high or low frequency of the target syntactic features were closely examined. The aim of this more fine-grained analysis was to identify how these features manifested themselves in the context of the items and whether they posed a threat to accessibility. Two such items are presented and discussed below. The items, which were of similar length (Item 1: 46 words; Item 2: 49 words), were obtained from Biology examination papers. The frequency of the target syntactic features for Item 1 and Item 2 can be found in Tables 6 and 7 respectively.

Item 1:

Hormonal control of […] is achieved by hormones acting on the […]. Using your knowledge of the way in which […] is coordinated, suggest why it can be deduced that hormones act on the […] rather than on individual […] cells.

Table 6: Item 1: Target syntactic features per 100 tokens (as generated by MAT)

Item 1						
Tokens	*Additional clauses*	*Passives*	*Prepositional adjuncts*	*"to" infinitives*	*Participles*	*Relative clauses*
46	0	4.35	17.39	0	4.34	4.34

Item 2:

The table below shows the mean […] rate and the standard deviation (SD) for the […] treatment group and the control group. Plot a suitable graph to show all the data for the […] treatment group. Do not include the standard deviations. Join the points with ruled, straight lines.

Table 7: Item 2: Target syntactic features per 100 tokens (as generated by MAT)

Item 2						
Tokens	*Additional clauses*	*Passives*	*Prepositional adjuncts*	*"to" infinitives*	*Participles*	*Relative clauses*
49	0	0	6.12	2.04	0	0

Item 1 can be described as a more complex text. The second sentence contains a present participle ("*Using*…") that modifies the main command verb "*suggest*", instructing students on what to do to answer the question. In addition, there are two nested clauses ("…*the way in which*…" and "…*why it can be deduced*…") and two passive structures ("…*is achieved by*…" and "…*is coordinated*…") which amount to 31 words, as well as multiple entities that need to be processed by the candidate.

In contrast, Item 2 comprises four short sentences, none exceeding 21 words, with simple subject-verb-object (sentence 1) and imperative-object structures (sentences 2, 3 and 4). The four sentences have mostly short noun phrases, contain minimal extra information in the form of prepositional adjuncts and no nested relative clauses. Also, there is no preceding modification of the main command verbs "*plot*" and "*join*".

Overall, of these two similarly sized items, Item 1 appears less accessible due to its longer sentences, its greater number of nested structures and its lengthy, left-branching participle leading up to the main verb in the second sentence ("*Using your knowledge of*…, *suggest*…" which requires the candidate to process additional linguistic material before reaching the main verb of the sentence).

Discussion

This study compiled three corpora of examination papers and used corpus linguistics techniques to explore linguistic accessibility in examination questions. The lexical and syntactic analyses to which the corpora were subjected, via AntWordprofiler and MAT respectively identified trends that invite closer attention.

AntWordProfiler, when used in conjunction with the vocabulary level lists, can help to identify low-frequency vocabulary that may inhibit reading comprehension, especially for candidates who do not have English as a first language. Vocabulary which does not represent content-obligatory language but is categorised above level 4 (i.e., it is at "proficient level" according to CEFR) might be considered complex and likely to introduce construct-irrelevant variance into test scores. When such vocabulary is identified by software and judged by question writers to be indeed complex, alternatives should be sought. A comparison of synonyms against the vocabulary level lists could help question writers to identify more accessible lexical substitutes. For instance, in the examples above, 'obstructed' could be used in place of 'hindered' (Level 5 vocabulary), while 'unwillingness' or 'hesitation' could be used in place of 'reluctance' (Level 5 vocabulary). Even though some words appear less sophisticated and therefore more accessible than others, it would be useful for future research to attempt to evaluate the effect of lexical substitutions on candidates' performance. Such evaluations may help to provide not only a more empirical basis for the need to exhibit lexical sensitivity in item writing but also indicate the forms that such lexical sensitivity should take in practice.

With respect to syntactic complexity, software such as MAT can be used to profile syntactically individual items and identify the frequency of features that could influence syntactic complexity. The qualitative comparison of pairs of more and less syntactically complex items of similar length may help to identify linguistic structures and item writing styles likely to prove barriers to accessibility. As shown in this study, examples of such linguistic structures and/or styles include left-branching constructions (signalled by features such as participles), the presence of multiple entities to be processed (signalled by features such as prepositional adjuncts), longer sentences (signalled by features such as additional clauses and prepositional adjuncts), and multiple and nested clauses (signalled by features such as relative pronouns and subordinating conjunctions). Where relatively inordinately high levels of such features are found in items, the items could be flagged for further

consideration and potentially for revision to improve accessibility. As the examples examined in this study indicated, items that displayed a higher concentration of these features appeared to be less accessible than similarly sized items that displayed a lower concentration of the target features. However, to enable the automated identification of excessively complex items in the future, further research is required. Such research can draw on developments in the field of linguistics and test in an experimental manner the accessibility of different linguistic configurations of items to help identify empirically-derived principles of linguistic accessibility.

In conclusion, corpus linguistics tools have not been typically used in item writing. However, as this study has demonstrated, they can prove particularly useful by providing directions for the improvement of items. Apart from helping to identify items that may display low levels of linguistic accessibility, they can also be used as training instruments in professional development courses intended for prospective as well as experienced item writers. Arguably, corpus linguistics tools can help to raise awareness among item writers of the ways in which different linguistic features and different item writing styles can hinder or enable the measurement of students' true abilities.

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.

Anthony, L. (2013). AntWordProfiler (Version 1.4.0w) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from: http://www.laurenceanthony.net/software

Anthony, L. (2018). AntConc (Version 3.5.2) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from: http://www.laurenceanthony.net/software

Ariel, M. (2001). Accessibility theory: an overview. In T. Sanders, J. Schilperood & W. Spooren (Eds), *Text representation: linguistic and psycholinguistic aspects* (pp.29–87). Amsterdam: John Benjamins.

Bauer, L., & Nation, I.S.P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253–279.

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2013). Test accessibility: item reviews and lessons learned from four state assessments. *Education Research International, 2013*.

Beglar, D., & Nation, I.S.P. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Biber, D., Douglas, B., Conrad, S., & Reppen, R. (2004). *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

Cloud, N., Genesee, F., & Hamayan, E. (2000). *Dual language instruction: a handbook for enriched education*. Boston, MA: Heinle & Heinle.

Council of Europe (2018). *Use of the CEFR*. Retrieved from: https://www.coe.int/en/web/common-european-framework-reference-languages/uses-and-objectives.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010, August). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp.276–284). Association for Computational Linguistics.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1–76.

Lenzner, T. (2014). Are readability formulas valid tools for assessing survey question difficulty? *Sociological Methods and Research, 43*(4), 677–698.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78*(2), 333–368.

Miller, G. A., & Isard, S. (1964). Free recall of self-embedded English sentences. *Information and Control, 7*(3), 292–303.

Nation, I.S.P. (2018). *The BNC/COCA word family lists*. Retrieved from: https://www.victoria.ac.nz/__data/assets/pdf_file/0004/1689349/Information-on-the-BNC_COCA-word-family-lists-20180705.pdf

Nini, A. (2015). Multidimensional Analysis Tagger (Version 1.3). Retrieved from: http://www.academia.edu/4285869/Multidimensional_Analysis_Tagger_v_1.3

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105–126.

Shaw, S., & Imam, H. (2013). Assessment of international students through the medium of English: Ensuring validity and fairness in content-based examinations. *Language Assessment Quarterly, 10*(4), 452–475.

Štajner, S., Evans, R., Orasan, C., & Mitkov, R. (2012). What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility* (pp.14–22).

Webb, S., & Nation, I. S. P. (2008). Evaluating the vocabulary load of written text. *TESOLANZ Journal*, 16, 1–10.

Wolf, M. K., Herman, J. L., Kim, J., Abedi, J., Leon, S., Griffin, N., Bachman, P. L., Chang, S. M, Farnsworth, T., Jung, H., Nollner, J., & Shin, H. W. (2008). Providing validity evidence to improve the assessment of English language learners. CRESST Report 738. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

A framework for describing comparability between alternative assessments

Stuart Shaw Cambridge Assessment International Education, **Victoria Crisp** Research Division and **Sarah Hughes** Cambridge Assessment International Education

Introduction

The credibility of an awarding organisation is partly reliant upon the claims it makes about its assessments[1] and on the evidence it can provide to support such claims. Some such claims relate to comparability. For example, for syllabuses with options, such as the choice to conduct coursework or to take an alternative exam testing similar skills, there is a claim that overall candidates' results are comparable regardless of the choice made. This article describes the development and piloting of a framework that can be used, concurrently or retrospectively, to evaluate the comparability between different assessments that act as alternatives. The framework is structured around four types of assessment standards and is accompanied by a recording form for capturing declared comparability intentions and for evaluating how well these intentions have been achieved. The framework and recording form together are intended to:

- provide a structure for considering comparability in terms of four established assessment standards;
- afford an opportunity for test developers to consider their intentions with respect to the comparability claims they wish to make;
- provide a list of factors (within each assessment standard) that are likely to contribute to the comparability of two alternative assessments;
- give a structure for collecting a body of relevant information against these factors;
- prompt an evaluation (on the part of the test developer) of how effectively the claims have been met.

Developing the comparability framework

Concepts of comparability and standards

This work focused on the comparability of assessment standards – in other words, the application of the same standard across different assessments (Newton, 2007). However, what is meant by *assessment standards* requires specification. In an attempt to explore assessment standards, we reviewed the relevant literature, with a focus on standards as associated with comparability in examination systems similar to our own. The search involved scrutinising the literature generated by awarding organisations both within the UK and internationally and the general comparability literature. Drawing on this literature, we identified four types of standard for the purposes of this work: content, demand, marking and awarding.

1. Note that the terms 'assessment' and 'test' are used interchangeably in this article.

- **Content standards** are about the value or relevance of the content of the assessment (Cambridge Assessment, 2010). They involve the appropriateness and coverage of the content specified to be assessed. They are also affected by how appropriate the specification or assessment criteria are and how well the questions are aligned to these. In addition, how well an assessment samples the content set out in the specification/syllabus is part of the 'content standards'.
- **Demand standards** are about the nature of knowledge, skills and understanding (KSU) required to successfully complete an assessment (Newton, 2005). This is evidenced in the degree of challenge in the questions and also relates to the level of accessibility of the assessment. The degree of challenge will be affected by the cognitive process(es) that students need to use to tackle the question. These are impacted on by the tools involved (e.g., paper, pencil/pen, notepad, calculator, ruler, computer, keyboard/mouse, computer screen, on-screen tools, response space on-screen) and the cognitive abilities of the candidate needed to answer the question. The tools provided will influence students' performance and experience of the assessment. For example, student familiarity with the tools they will need to use during their assessment (e.g., with the software platform used in an on-screen assessment) is likely to influence performance.

  Content and demand standards are related but the distinction between them is useful. The content standard relates to the appropriateness and coverage of topics, whereas the demand standard relates to what the student is expected to do in relation to the topics. However, we recognise that there may be some overlap in terms of the topic and the demands it makes on the student.
- **Marking standards** are about how marks are assigned to reward the knowledge, understanding and skills shown in students' performances. Marking standards also relate to the degree of leniency or severity of marking (Pinot de Moira, Massey, Baird, & Morrissy, 2002; Cambridge Assessment International Education, 2017). Marking standards are inherent in the mark scheme, where the underlying knowledge, understanding and skills to be rewarded are defined. Marking standards are also affected by the marking processes, the compliance of marking processes with codes of practice, the accuracy of the marking, the competence of examiners and adequacy of any standardisation or moderation procedures. This type of standard relates to how well scores reflect the constructs that the assessment is intended to measure.
- **Awarding standards** are about the results that students achieve on an assessment (the assessment outcome, e.g., a grade) and about the kinds of performances that should receive a particular outcome

(Coe, 2010; Baird, Cresswell, & Newton, 2000). In other words, these standards are about the scores that will receive a particular grade. When grading the assessment, the aim will (almost always) be to maintain the awarding standard applied in previous sessions. Awarding standards are affected by the procedures and policies in place to support grading and by the combination of technical and statistical evidence and professional judgement used in order to determine cut scores.

It is both possible and reasonable for a pair of assessments to be comparable in terms of awarding standards but to not be comparable in terms of content standards, demand standards, or marking standards. For example, two optional assessments within a syllabus might test different topics, different demands, and be marked against a different marking scale but can be considered comparable in terms of awarding standards (though not in terms of the other standards) if the grading process ensures that the same grades are given for equally competent performances.

In deciding on this framework, we have gone beyond the traditional structure of content standards (defining what students should learn) and performance standards (the evidence types needed to demonstrate the content and the quality of student performance that is considered worth a particular grade) (see Linn, 1994). We are using content standards to refer to the content assessed (which will be a subset of the content to be learnt), as this is important in evaluating the comparability of two assessments that are alternatives within a qualification. Additionally, we have replaced performance standards with demand, marking and awarding standards. This provides a more detailed framework for use to support comparison between assessments.

Building the comparability framework and recording form based on the four comparability standards

The purpose of the comparability framework is to outline the criteria for comparability for the four types of standard described. The framework comprises four columns representing the four assessment standards. The ordering of the standards reflects their influence at different stages throughout the test design and testing process. Each standard is fronted by a conditional statement, for example: "If it is the intention that content standards are comparable across assessments, the following need to be fulfilled." What then follows is a list of factors that need to be the same across alternative assessments for there to be comparability with regard to that standard. By way of illustration:

- In the case of *demand* standards, one of the listed factors states that the range of kinds of questions or tasks should be the same across assessments. For example, there should be a similar balance of question types (e.g., MCQ, short answer, essay) on each of the assessments compared.
- In the case of *marking* standards, one of the listed factors states that the application of the mark scheme should be the same across assessments with markers complying with marking guidance and requirements for both assessments.

The comparability recording form provides opportunities to identify which comparability standards are intended as claims, space to record any differences between assessments for each of the standards, and an opportunity for making an overall judgement.

The intentions are likely to depend on the purpose of the assessment in terms of how it relates to the qualification as a whole. If there is no intention for there to be comparability with regard to a particular standard then the relevant rows can be ignored. Where differences are identified, then any efforts made to address them can be recorded. Differences suggest potential threats to comparability. By addressing such threats, comparability between assessments can potentially be achieved. For example, in the case of comparing an on-screen and a paper-based assessment, if a certain skill cannot be assessed directly on-screen, efforts might be made to provide functionality that allows candidates to show their skills in this area in a comparable way.

Ultimately, it is necessary to determine whether comparability is achieved for each of the standards where it is intended. Whilst all differences are potential threats to comparability, it may be that not all of them are serious threats, and some threats may have been mitigated by efforts to address them (as recorded in the form). This is a judgement that needs to be made in light of the context of the qualification. For example, the omission of a particular subtopic on one of two alternative assessments might have a more or less serious effect on comparability depending on how important the subtopic is within the syllabus. Given the ways in which the differences are addressed, a judgement is necessary as to whether comparability between assessments is sufficient for them to be considered comparable alternatives within the same qualification for each of the standards where comparability is intended.

Piloting the comparability framework

We wished to explore whether those involved in creating assessments could use the framework and form in the way we intended, and whether they found it helpful. To do so, we conducted the pilot exercise described below.

Assessment contexts

The framework and recording form were piloted using two Cambridge Assessment International Education assessment contexts where there are two assessments that act as alternatives:

- On-screen and paper-based tests: Stage 8 Progression tests[2] in Science for 2018, Papers 1 and 2 (both available as either on-screen or paper-based).
- An Alternative to Practical exam paper and a Practical test: IGCSE[3] Chemistry (0620, for June 2017).

Assessment materials

Materials specific to the relevant assessment context were used in the piloting. These were:

- For IGCSE Chemistry
  - IGCSE Chemistry (0620) Syllabus;
  - IGCSE Chemistry Practical test (June 2017), instructions and mark scheme;
  - IGCSE Chemistry Alternative to Practical (June 2017) and mark scheme.

2. Cambridge Primary/Lower Secondary Progression Tests are end-of-stage tests which are designed to measure learners' progress and identify their strengths and weaknesses.
3. The Cambridge International General Certificate of Secondary Education (IGCSE) is a general education qualification for 14 to 16 year olds, available in a range of subjects.

- Stage 8 Science Progression tests
  - Cambridge Lower Secondary Science Curriculum;
  - Paper-based Stage 8 Progression tests (2018, Papers 1 and 2) and mark scheme;
  - Links to the on-screen versions of the Stage 8 Progression tests (2018, Papers 1 and 2).

Participants and procedure

For each assessment context, an expert was recruited who was known to have a familiarity with, and expertise in, the selected context (in terms of their setting and marking experience). Both experts were asked to:

- Read a report that was provided in order to familiarise themselves with the comparability framework, recording form and guidance on how it was envisaged these could be used.
- Re-familiarise themselves with the target assessment materials (as provided).
- Complete the comparability recording form that accompanies the framework with appropriate details. Participants were asked to refer to the assessment materials themselves and to use their knowledge of how the assessments were created, marked and graded. (If there were parts of the process with which they had no or little experience for these assessments, they were asked to leave the relevant boxes blank.)
- Complete a questionnaire in order to provide feedback on use of the framework and recording form, including thoughts on how it could be used in the future.

Feedback from participants

Feedback from the two experts on use of the framework and form, as provided in their questionnaire responses, are summarised below in terms of salient themes. The feedback led to changes to the framework and form. The reader may find it useful to refer to Tables 1 and 2, which show the revised framework and form, when specific points within them are mentioned in this section.

Comprehensibility of the comparability framework and recording form:

- In the main, the framework was largely understood by participants. This was aided by reading the report about the framework. Similar preliminary reading would be required for future users who might be unfamiliar with some of its concepts (e.g., the different 'standards').
- The standard relating to 'demand' proved to be the most challenging to comprehend. Despite the challenges, one participant acknowledged that the 'demand' standard has the potential to extend the thinking of the user beyond merely 'content' comparison and encourage consideration of the cognitive processes students need to employ to tackle a question.
- The bulleted points, 'cognitive processes' and 'range of kinds of questions' (both relating to demand standards) were undoubtedly thought-provoking to the participants. Their comments suggest that these themes required the most thought as they reflect key differences between alternative assessments that are difficult to avoid (e.g., risk of inaccurate results during a practical, effects of working on-screen on the cognitive processes used).
- There was some perceived overlap between certain bullet points within the comparability framework. This suggests an inability on the part of the user to reliably distinguish between concepts in the framework. However, perceived areas of commonality may be more attributable to a lack of understanding of lexical terms ('domain' and 'topic', for example) than to truly indistinguishable categories. (The perceived areas of overlap have been addressed with revisions to wording, see later.)
- One participant reported challenges in how to use two of the columns within the comparability framework. In the draft of the framework used in the pilot, column 4 asked for differences between the two alternative assessments to be recorded and column 5 was to be used to record how these differences had been addressed. The participant felt that in terms of the paper production and marking processes "the differences have already been addressed as far as possible – and so to identify a difference and then state how they have been addressed is difficult." The participant made some suggestions for revisions to these columns.

Usefulness and usability of the comparability framework and recording form:

- The framework and form were considered useful by the participants especially in terms of providing criteria for assessing comparability between tests that will be treated as equivalent. For example, the criteria should ensure that the focus does not rely too heavily on test content without considering other elements of comparability. In addition, the importance of having the same senior examiners (or at least an overlap of senior examiners) involved in marking two optional tests is reinforced by the completion process (as is the need to maintain question similarity across test forms). However, some features of the framework and elements within the recording form were deemed to be beyond the control of the participants (such as standardisation methods and quality assurance). This is not necessarily problematic in terms of it being possible for users to complete the form but emphasises that some users may be better placed than others to address certain differences between assessments in order to improve comparability.
- Making a judgement in the final column in the recording form (*For the standards where comparability is intended, are you satisfied that there is sufficient comparability?*) appeared to present minimal problems to participants. However, despite differences acknowledged in other columns in the form, participants answered 'yes' in the final column for all of the standards. Given their extensive involvement in the qualifications, it is possible that participant's responses may be somewhat skewed. Alternatively, it may be that the differences are genuinely seen as trivial and not thought to compromise comparability.
- The participants felt that the completed recording form can provide evidence to support the stated (intentional) claims of comparability made by the test developer.
- Participant comments suggested that using the comparability framework and completing the recording form did not provide new insights for those involved in the qualifications. However, as mentioned earlier, participants reported that it provides a set of criteria for considering comparability issues and avoids certain concerns being over- or under-emphasised. Therefore, using the form to systematically consider and record information relating to comparability will still be valuable.

- As a tool for retrospectively evaluating the comparability of two alternative assessments, both participants considered the framework and form valuable (though this perception was subsequently caveated by one participant who argued that any form of retrospective analysis might be considered somewhat tardy).
- Participants felt that the framework could be used beyond the contexts in which it was piloted, wherever parallel routes to certification exist.
- Participants reported that the framework would be useful throughout the test development process but would be most helpful at the setting stage. The framework and recording form might be used at different times during test construction and by different assessment personnel.
- Participants felt that the recording form could constitute an additional source of comparability evidence (alongside existing evidence such as specification grids and statistical data), providing that completion of recording forms does not degenerate into a mechanical checklist exercise.

Frequency of application of the comparability framework and recording form:

- The application of the framework and recording form is not perceived as being necessary every time that parallel assessments are created and used.
- Participants felt that the comparability framework and recording form could be useful when syllabuses are reviewed, and the first time an alternative assessment is created (to parallel an existing assessment).

Users of the comparability framework and recording form:

- One participant reported that the comparability framework and recording form should be thought of as "an organic document that is amended and changed during the life-time of the test."
- Participant responses suggested that a range of personnel with roles within the (re)development of an assessment should be engaged in using the comparability framework and form at different stages throughout the assessment process, for example:
    - Revisers[4]: could be tasked at the revising stage with some responsibility for completing parts of the form when checking for comparability;
    - QPEC (Question Paper Evaluation Committee) personnel: could complete parts of the form when reviewing the assessment materials;
    - Principal Examiners[5]: could use the form when considering what grade thresholds to recommend to the grading team;
    - Assessment Managers[6]: should have a responsibility for declaring the intended comparability claims (i.e., whether the assessments are intended to be comparable with regard to each of content, demands, marking and awarding) and for final evaluation of whether there is sufficient comparability for each dimension where comparability was intended.

4. After questions have been drafted by a setter, revisers provide constructive, expert feedback, checking that the question paper and mark scheme match the syllabus, contain accurate content, are of appropriate demand, and avoid construct-irrelevant effects.
5. Principal Examiners oversee the marking of student responses and are responsible for standards in the marking of examination scripts.
6. Assessment Managers oversee all stages of the creation and use of the assessments for a particular syllabus. They are responsible for standards in a particular examination and over time.

Changes to the structure and content of the recording form

As mentioned earlier, feedback from the pilot participants led to some revisions to the comparability framework and recording form, as detailed below. The revised framework and form are shown in Tables 1 and 2.

- There was potential overlap (giving rise to possible ambiguity) between 'topic' and 'subtopic' in the content standards section of the original form. These two categories were conflated into one row as simply 'subject topics'.
- Reference to *Assessment Objectives* was added within the point relating to 'knowledge, understanding and skills' in the demand standards (this already appeared in the framework itself).
- Reference to 'range of kinds of questions' in demand standards was changed to 'range of kinds of questions/tasks' to ensure this point encompasses a wide variety of assessment task types.
- Due to potential overlap, the categories 'standardisation methods' and 'quality assurance processes' in marking standards were merged into one row to read 'standardisation methods and any other quality assurance processes'.
- One issue that was raised during the pilot was how, when conducting a retrospective comparability evaluation, it was difficult to identify differences between tests (original column 4, *What are the differences between tests, if any, in terms of these features?*) and then how those differences had been addressed (original column 5, *How have the differences been addressed (if they have been)?*), as differences that had been addressed might not be observable in the final materials. As a consequence, the original column 5 was removed. Column 4 was retained and a note was added that actions to minimise differences could also be recorded in this column. This should allow the form to be appropriate for both concurrent and retrospective evaluations.

Conclusions

The comparability framework constitutes a structure for considering four comparability standards when developing an alternate assessment. The comparability recording form affords a means for capturing comparability intentions and for evaluating whether those intentions have been achieved.

A number of issues emerged from both the developmental work on the framework and subsequent piloting:

- The value and application of the framework and recording form should extend beyond the two kinds of contexts with which they were piloted (paper-based and computer-based comparisons and Alternative to Practical and Practical tests) and may include a number of other contexts where there are optional assessments within a qualification.
- There are a number of circumstances in which an evaluation of the comparability of parallel routes might be desirable. For example, where a new assessment is being introduced as a parallel to an existing assessment; where the comparability of two alternative assessments has been queried; or where a qualification containing parallel optional assessments is undergoing routine review with a view to redevelopment. There are two options for how a comparability review using the framework and form can be

Table 1: Revised version of comparability framework

Comparability of:			
Content standards	*Demand standards*	*Marking standards*	*Awarding standards*
If it is the intention that content standards are comparable across tests, the following need to be fulfilled: • subject domains are the same across tests; • subject topics are the same across tests; • whole test content coverage is the same across tests.	If it is the intention that demand standards are comparable across tests, the following need to be fulfilled: • knowledge, understanding and skills (e.g., Assessment Objectives) assessed are the same across tests; • the range of kinds of questions or tasks are the same across tests (e.g., similar balance of MCQ, short answer, essay); • the test environment does not affect the nature of the teaching and learning; • the test environment is easy to use and students have been given sufficient opportunity for familiarisation with the test environment; • the cognitive processes (as supported by tools) are the same across tests as far as we can tell; • the possible effects of any differences in response format are carefully considered (e.g., for on-screen tests, the effects of typing rather than writing on paper, or of using a drop-down list rather than circling a response on paper).	If it is the intention that marking standards are comparable across tests, the following need to be fulfilled: • the mark schemes reward the same knowledge, skills and understanding; • the application of the mark scheme is the same across tests with markers complying with marking guidance and requirements for both tests; • the way that student responses are presented to markers needs to give equal opportunity for accurate marking across tests; • marker competence/accuracy is the same across tests (ideally, the same specific markers are used for both tests); • markers are standardised appropriately for both tests and appropriate quality assurance processes are used for both tests; • auto-marking (if used) and human marking are both sufficiently accurate and reward intended constructs (only relevant if comparing an on-screen test to a paper-based test).	If it is the intention that awarding standards are comparable across tests, the following need to be fulfilled: • awarding is conducted separately for different tests with potentially different grade thresholds (thus ensuring comparability of awarding standards between tests even if there are differences in content, demand or marking standard); • the awarding process is the same across tests (e.g., use of judgemental and statistical evidence, methods of recording awarding decisions); • sufficient data is available to compare across tests (e.g., entry sizes, benchmark centres, syllabus pairs, knowledge of the characteristics of the candidates entering for each test); • awarding standards are maintained over time across tests.

Table 2: Revised version of comparability recording form

Comparability recording form: a structure for describing comparability across tests

Completed by (name)…….....…… (job role)…… Date……

Assessment name and code……...

1. Standard	*2. Is it intended that there should be comparability between tests in terms of each standard?*	*3. Comparability features – these should be the same across tests if comparability between tests is intended for that standard*	*4. What are the differences between tests, if any, in terms of these features? (Notes can also be included on actions taken to minimise differences)*	*5. For the standards where comparability is intended, are you satisfied that there is sufficient comparability?*
Content standards		Subject domains*		
		Subject topics*		
		Whole test coverage		
Demand standards		Knowledge, understanding and skills (e.g., Assessment Objectives)		
		Range of kinds of questions/tasks		
		Teaching and learning		
		Test environment ease of use and opportunity for familiarisation		
		Cognitive processes		
		Response format		
Marking standards		Mark schemes		
		Application of the mark scheme		
		The way that student responses are presented to markers		
		Marker competence/accuracy		
		Standardisation methods and any other quality assurance processes		
		Any auto-marking is sufficiently accurate and rewards intended constructs (only relevant if comparing an on-screen test to a paper-based test)		
Awarding standards		Awarding conducted separately for different modes		
		Awarding process		
		Sufficient data is available		
		Awarding standards are maintained over time		

*For example, for a Physics assessment subject domain would refer to areas such as 'electricity and magnetism' and subject topic to aspects such as 'electric circuits'.

conducted, the choice of which will be influenced by the circumstances of the evaluation. The two options are:

  - *Concurrent* – During the development of the assessments for a particular examination session (i.e., a particular administration of the assessment), those involved use the comparability framework and form at intervals to guide aspects of the assessment design and to monitor success in achieving comparability. The form can be updated alongside the papers' development, administration, marking and grading, thus providing an audit trail and record of efforts made to achieve comparability.
  - *Retrospective* – After the development, administration, marking and grading of the assessments for a particular session, those who were involved use the framework and recording form to review the comparability of the tests based on relevant documents and their own experience of involvement in parts of the process.

- The inherent value of the form is in its potential to capture substantive qualitative features of comparison (and not simply a checklist set of yes/no responses). Therefore, thoughtful consideration of the assessments needs to be encouraged when the framework and form are used.
- There is enough evidence from the pilot and preliminary (albeit tentative) evaluations that the comparability process provided by the framework and recording form could be used to enhance the professional development of examiners, conveying as it does the need to consider and apply several comparability standards.
- Information marshalled in support (or otherwise) of 'content', 'demand' and 'marking' standards might inform the awarding process.
- The Assessment Manager (person responsible for the assessment) is likely to be best placed to have overall responsibility for a comparability evaluation, beginning the process of form completion themselves and then passing the form to other relevant personnel as needed. Whilst there was some variation in the personnel that our pilot participants suggested as appropriate to complete each part of the framework, and there might sometimes be reasons for varying who is involved, some commonalities emerged allowing us to suggest the general pattern in Table 3 (note that the suggestions given here are specific to Cambridge Assessment International Education and may not necessarily generalise to other awarding bodies).

The comparability process outlined here affords a greater level of granularity of reporting for awarding bodies when making comparability claims regarding alternate options within the same syllabus. Not only can claims of comparability be made at a general level (qualification and subject), they can be made in light of specific standards of comparability making clear to stakeholders which of the four assessment standards are applicable. Importantly, standards for which comparability cannot be claimed (intentionally or otherwise) can be identified and described in greater detail than is currently reported.

The framework and form provide a tool that can be used to evaluate the comparability claims made regarding alternative assessments. The resulting evidence may provide support to the argument for the comparability of the parallel tests or provide insights that can inform adjustments to ensure comparability. Whilst the development and piloting of this tool has focused on general qualification contexts, the comparability framework and form might equally be applicable to vocational and technical qualifications.

References

Baird, J., Cresswell, M., & Newton, P.E. (2000). Would the real gold standard please step forward? *Research Papers in Education, 15*(2), 213–29.

Cambridge Assessment. (2010). *Exam standards: the big debate – Report and recommendations*. Cambridge: Cambridge Assessment. Retrieved from: http://www.cambridgeassessment.org.uk/Images/125765-exam-standards-report-and-recommendations.pdf

Cambridge Assessment International Education. (2017). *Code of practice.* Cambridge: CAIE. Retrieved from: http://www.cambridgeinternational.org/images/416992-code-of-practice.pdf

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education, 25*(3), 271–284.

Linn, R.L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4–14.

Newton, P.E. (2005). Examination standards and the limits of linking. *Assessment in Education: Principles, Policy and Practice, 12*(2), 105–23.

Newton, P.E. (2007). Contextualising the comparability of examination standards. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Pinot de Moira, A., Massey, C., Baird, J.A., & Morrissy, M. (2002). Marking consistency over time. *Research in Education, 67*(1), 79–87.

Table 3: Proposal for appropriate personnel to complete the comparability recording form*

1. Standard	*2. Is it intended that there should be comparability between tests in terms of each standard?*	*4. What are the differences between tests, if any, in terms of these features? (Notes can also be included on actions taken to minimise differences)*	*5. For the standards where comparability is intended, are you satisfied that there is sufficient comparability?*
Content standards	Assessment Manager	Question Setter[7] and Reviser	Assessment Manager
Demand standards	Assessment Manager	Question Setter and Reviser	Assessment Manager
Marking standards	Assessment Manager	Principal Examiner	Assessment Manager
Awarding standards	Assessment Manager	Principal Examiner and awarding team	Assessment Manager

*Note that column numbers match those in Table 2.

7. Question Setters set and develop a draft paper and mark scheme, paying attention to matching the syllabus, accuracy of content, appropriateness of demand, and avoidance of construct-irrelevant effects.

Comparing small-sample equating with Angoff judgement for linking cut-scores on two tests

Tom Bramley Research Division

Introduction[1]

The educational measurement literature makes a clear distinction between the activities of standard setting on the one hand, and test equating and linking on the other. For example, these topics occupy different chapters in the standard reference work Educational Measurement (Brennan, 2006). Test equating is usually defined in a fairly narrow, technical way such as: "Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably" (Kolen & Brennan, 2004, p.2). Standard setting, on the other hand, is usually defined more broadly such as "…the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states of performance" (Cizek, 1993, p.100). The main issues in test equating tend to be around the definition of the 'correct' equating transformation, and the data collection designs and statistical methods necessary to estimate it. In standard setting, however, the procedures are "… seldom, if ever, impartial psychometric activities, conducted in isolation. Social, political and economic forces impinge on the standard-setting process…" (Cizek & Earnest, 2015, p.213). In particular, standard setting processes involve human values and judgements, and differences in these are to be expected.

Conceptually, however, the processes of standard setting and test equating are clearly very closely related. The performance standard can be conceived of as a point on an abstract continuum, and the aim of the standard setting process as being to find the score on the raw scale of the particular test at hand that corresponds to this point. This seems very similar to the conceptualisation of equating in Item Response Theory (IRT) – the raw scores on two tests that correspond to the same level on the unobservable underlying trait are deemed equivalent.

If we are prepared to conceive of the abstract continuum on which the performance standard is located and the latent trait of the IRT model as one and the same, then we can see that carrying out separate standard setting exercises on Tests X and Y is in theory no different from attempting to equate them (at the point on the latent trait corresponding to the cut-score) by an IRT approach. Of course, the *results* of applying such dramatically different approaches to the same problem could be expected to differ.

Although it would seem most logically justifiable to carry out a standard setting exercise just once (to establish one definitive example of a realisation of the abstract performance standard on a concrete test) and then to use statistical equating to link all subsequent (or other) forms to that, in practice it may well be that a standard setting method is used (perhaps alongside other methods) to inform or set the cut-score on subsequent forms. Thus, the standard setting method is used in practice as a test equating (standard maintaining) method. There are several scenarios where this might arise, for example:

i) if the test is very high-stakes (e.g., a licence-to-practise test) where procedures require 'stakeholder' involvement in setting the cut-score on each test form;

ii) if sample sizes are so low on each test form that statistical equating methods are not trusted;

iii) if contextual factors (such as cost, need for test security, local culture and expectations) prevent some of the necessities for equating methods such as pre-testing, administration of an anchor test, or embedding of field-test items into live tests;

iv) if there is a need to determine a cut-score before any 'live' performance data has been collected.

The conceptual similarity between equating and standard setting raises questions of the relative accuracy[2] of the two methods. Our starting assumption was that in an ideal world a large-sample equating exercise would be the preferred way to map a cut-score from one test to another parallel one. However, since the standard error of equating in a test equating exercise depends upon the sample size, continually reducing the sample size presumably will reach a point at which the equating error becomes greater than the error that would arise from carrying out two separate standard setting exercises. The equating error from the latter will depend on the details of the method used, but for all methods that rely on the judgement of item difficulty by experts, a fundamental issue is the extent to which those judgements correspond to the actual empirical difficulty. One of the motivations for this research was the realisation (see Benton, 2020, this issue) that estimates of item difficulty based on extremely small samples of empirical data (N<10) can correlate better with the actual (full population) values than estimates based on expert judgement. The aim of this study was to compare, by simulation, the accuracy of mapping a cut-score from one test to another by expert judgement versus the accuracy with a small-sample equating method.

Method

Standard setting method

The standard setting method we simulated was the 'mean estimation' method – a variant of the more well-known Angoff Method (e.g., Loomis

1. This is a shortened and simplified version of a paper presented at the AEA-Europe conference in 2017 (Bramley & Benton 2017).

2. In this article we use 'accuracy' in the general sense of overall accuracy including both bias and random error.

& Bourque, 2001). It is applicable to tests containing polytomous as well as dichotomous items. If the test consists solely of dichotomous items it is the same as the Angoff Method. Experts estimate the difficulty of each of the items in a test in terms of the mean score likely to be obtained on each item by a group of minimally competent examinees (MCEs). If the test is pass-fail then the MCEs are those who are just competent enough to pass. If the test is graded into more than two categories, there are different groups of MCEs for each cut-score. The cut-score is derived by summing the estimated means and then averaging across judges, rounding the result to an integer if necessary (or averaging and then summing – it makes no difference).

Previous research (e.g., Impara & Plake, 1998) has suggested that although estimating the mean scores of MCEs can be difficult for experts in an absolute sense, they are more adept at discerning the correct rank order of the difficulty of items. Hence, judgements from experts can potentially be transformed onto the correct scale before being used to inform standard setting (Thorndike, 1982; Humphry, Heldsinger, & Andrich, 2014). Since judgements can be transformed to the correct scale, the correlation between estimated difficulties and actual difficulties (often measured by item facilities – mean mark divided by maximum possible mark) provides a reasonable idea of the value of the information from such methods, as discussed above. In our simulation (described in more detail later) we wanted to vary this level of correlation and assess the effect on the outcome.

Equating method

There are a variety of equating methods appropriate for use with small samples (for example, see Livingston & Kim, 2009; or Kim, von Davier, & Haberman, 2008). We wanted a method suitable for the 'non-equivalent groups anchor test' (NEAT) design. This is because for equating test forms which are only produced once or twice a year (such as GCSEs or A Levels) it is not usually possible to get one group of examinees to take both forms, or to obtain randomly equivalent groups of examinees. It is much more frequently possible to obtain two different groups and adjust statistically for differences in ability between them by means of an anchor test. We chose chained linear equating (e.g., Puhan, 2010) because it requires fewer parameters to be estimated than the theoretically preferable (with large samples) equipercentile equating. Puhan (2010) reports that, across a range of conditions, chained linear equating tends to perform well compared to other linear equating techniques for the NEAT design.

We were also interested in exploring the effect of clustering on the small-sample equating outcome. In practice, it might only be logistically feasible to obtain examinees from a single class in a small number of schools for an equating exercise, so it was of interest to see how a small clustered sample differed from a genuinely random sample of the same size.

In brief, the equating scenario consisted of a Test X (where we assumed the cut-scores were known) and a Test Y where we needed to set equivalent cut-scores. We simulated mean estimation judgements at two levels of correlation (0.6 and 0.9) between estimated and empirical values, and derived the cut-score on Test Y by adding up the simulated means for the items on Test Y. We compared this with a chained linear equating method in two conditions:

1. Random samples of 30 examinees from three schools in Group A took Test X, and from three different schools in Group B took Test Y, and all 180 examinees took an anchor Test V;
2. Simple random samples of 90 examinees in Group A took Test X and in Group B took Test Y, and all 180 examinees took an anchor Test V.

In both cases we considered two cut-scores, one at the lower end of the raw score scale and one at the higher end.

Data

The dataset forming the basis of all the analyses reported here was artificially constructed from a large real dataset containing the responses of 15,731 examinees to a test with a maximum possible raw score of 200. The questions were made up of sub-questions (henceforth items), and the items ranged in tariff (maximum score) from 1 (i.e., dichotomous) to 5 (i.e., polytomous with six score categories). The facility values of all items were calculated and two Tests X and Y, each with a maximum possible raw score of 60, were constructed by selecting two sets of items comprising fifteen 2-tariff items and ten 3-tariff items by systematically alternating selection from the items ordered by facility value. An anchor Test V was constructed from 20 dichotomous items (which was all the dichotomous items and hence no selection method was required).

The examinees came from 323 schools, each contributing between 1 and 238 examinees (mean 48.7, median 33). Each school had a 5-digit identification number, which was known to be non-randomly assigned. Two non-equivalent groups of examinees of roughly the same size were created by assigning those in schools with ID numbers below a certain value to Group A and the rest to Group B. Scores on the anchor test correlated around 0.8 with scores on Test X and Y in both groups.

Table 1 shows that Test Y was slightly easier than Test X (higher mean score) but the lower SD of scores on Test Y shows that the difference in difficulty was not uniform across the score range. It is also clear that Group A was of higher ability than Group B (its mean score was higher on all tests).

Table 1: Descriptive statistics for scores on Tests X, Y and V

Test	*All (N=15,731)*		*Group A (N=7,752)*		*Group B (N=7,979)*	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
X (max 60)	31.76	13.29	33.00	13.39	30.55	13.08
Y (max 60)	32.36	12.16	33.46	12.37	31.29	11.86
V (max 20)	10.01	3.44	10.30	3.50	9.72	3.34

As described in the introduction, because of the conceptual similarity between the 'abstract continuum' on which the performance standard is located and the 'latent trait' of IRT, we defined the correct equating function to be the one arising from IRT true score equating on the complete dataset (i.e., X, Y and V items calibrated concurrently for both groups in a single-group design with no missing data). We focused on two different cut-scores on Test X: 15 out of 60, and 45 out of 60. The 'definitive' equated cut-scores on Test Y arising from the IRT true score equating were 17.20 and 44.21.

Equating via simulating judgements in a standard setting method

We simulated expert judgement of item difficulty by adding random error to the 'correct' (empirical) values. We simulated two levels of correlation: 0.6 (a value representative of published Angoff studies,

see for example Brandon, 2004), and 0.9 (a much higher value than usually found, in order to represent a very optimistic view of what might be achievable in ideal conditions).

The technical details of the simulation are described in Bramley and Benton (2017). The process was repeated 1,000 times for each of two different values of the correlation *r* (0.9 and 0.6) and for the two different Test X cut-scores (15 and 45).

The simulated judgements were used to produce equated cut-scores, using the standard setting method previously described. The distributions of equated cut-scores were then compared with the definitive (correct) cut-score. Specifically, bias B was defined as the mean difference (across replicates) between the equated score for each replicate and the correct cut-score; error variance E was defined as the variance of the equated cut-scores; and the root mean squared error RMSE (Root Mean Square Error) was calculated as sqrt(B^2+E).

Equating via a traditional small-sample equating method

For condition 1, all schools with 30 or more examinees were selected and then a two-stage sampling process first selected at random three schools from each group, and then a random sample of 30 examinees from each school. This process was replicated 1,000 times. For condition 2, we selected 1,000 simple random samples (with replacement) of 90 examinees from Group A and 90 from Group B. An equated cut-score on Test Y for each of the Test X cut-scores (15 and 45) was derived by chained linear equating in each replicate in each condition (see Bramley & Benton, 2017 for the equations). The distribution of equated scores across the 1,000 replicates was then compared with the definitive cut-score in the same way as for the simulated judgements.

Results

Table 2 shows that in all cases except small-sample equating with the clustered sample (condition 1) the bias made a negligible contribution to the overall RMSE. The more realistic value for the correlation (0.6) had RMSE values nearly twice as high as that for the optimistic value (0.9) at both cut-scores. The % distributions in Table 2 refer to equated cut-scores on Test Y rounded to the nearest integer. This is on the assumption that in practice, if an integer cut-score were required to be set on Test Y, the correct values would be 17 and 44. This causes a slight asymmetry because an equated score of 44.6 (say) would be rounded to 45 and be 1 too high, whereas a less accurate equated score of 43.6 would be rounded to the correct value of 44. For simulated correlations of 0.9, the equated cut-score was within ±1 of the correct score around 75% of the time (cut-score of 15) or 80% of the time (cut-score of 45), but for simulated correlations of 0.6 only around 50% were in this range, and around 25% were three or more score points away.

The overall accuracy of small-sample equating, as measured by the RMSE, was better in condition 2 (simple random sample of 90 examinees from each test) than in condition 1. At both cut-scores, the condition 2 RMSE was roughly half-way between the RMSE values from simulated judgements with r=0.6 and r=0.9. The condition 1 RMSEs were about 0.7 score points higher than the corresponding condition 2 RMSEs, for both cut scores, showing the detrimental effect of clustering of examinees within schools on equating error. The condition 1 RMSEs were slightly higher than those from simulated judgements with a correlation of 0.6. In the best case for small-sample equating (condition 2) the cut-scores were within one score point of the correct value around 60% of the time for a cut-score of 15 and around 70% of the time for a

Table 2: Equated scores based on simulated judgements and small-sample equating (replications=1,000)

	Simulated judgement		*Equating condition…*		*Simulated judgement*		*Equating condition…*	
	r=0.6	*r=0.9*	*1*	*2*	*r=0.6*	*r=0.9*	*1*	*2*
Test X cut-score	15				45			
Correct Y cut-score	17.20				44.21			
Test Y mean equated cut-score	17.08	17.15	16.39	16.56	44.38	44.33	44.25	44.36
Test Y SD equated cut-score	2.30	1.25	2.41	1.70	2.06	1.12	2.18	1.45
Bias	–0.12	–0.05	–0.81	–0.64	0.18	0.12	0.04	0.15
RMSE	2.31	1.25	2.54	1.82	2.07	1.13	2.18	1.45
%<= –3	12.1	0.9	19.1	12.0	8.8	0.9	10.5	1.9
% –2	13.3	8.1	10.4	13.6	9.3	4.3	9.9	7.9
% –1	16.4	21.9	18.8	22.1	14.3	17.6	16.9	17.5
% 0	17.6	29.6	19.7	23.2	18.4	32.0	19.0	27.1
% +1	15.0	25.6	14.1	16.0	19.4	31.0	16.2	24.6
% +2	10.3	11.3	10.3	9.7	13.8	12.0	11.4	13.4
% >= +3	15.3	2.6	7.6	3.4	16.0	2.2	16.1	7.6

cut-score of 45. Bias made a small contribution to the RMSE at a cut-score of 15 and a negligible contribution at a cut-score of 45. The fact that sampling error was the main contributor to RMSE in all methods and conditions suggests that comparisons are not critically dependent on how the 'true' equating function is defined, because this would only affect the bias and not the sampling error.

Discussion

This study has compared, by simulation, the level of accuracy that might be obtained from a standard setting method (mean estimation) if applied as a test equating method to that which might be expected from a small-sample test equating method (chained linear equating). As expected, the standard setting method resulted in more accurate equating when we assumed a higher level of correlation between simulated expert judgements of item difficulty and empirical difficulty. For small-sample equating with 90 examinees per test, more accurate equating arose from using simple random sampling compared to cluster sampling at a given sample size. The actual values of RMSE depended on the cut-score, being generally larger for the cut-score where the correct equated cut-score on Test Y was further from the cut-score on Test X. The simulations based on the more realistic value for the correlation between judged and empirical difficulty (0.6) produced a similar RMSE to small-sample equating with cluster sampling. Simulations of standard setting based on the optimistic correlation of 0.9 had the lowest RMSEs of all.

As shown by Benton (2020, this issue), even very small samples of examinees can give a more accurate picture of the relative difficulty of items than estimates from experts. We may therefore be surprised that the small-sample approach trialled here did not perform even better. There are a number of reasons for this. One reason is that the equating approach adopted in the simulation study required calibration of examinee abilities across two groups using an anchor test. Small-sample equating with a single group design would be significantly more accurate. Even within the NEAT design, it may be that other approaches, such as Tucker linear equating or Rasch true score equating, may provide a more stable estimate of equivalent scores than chained linear equating.

Most important, however, is the fact that our simulations assumed that judged and empirical values for the mean scores of MCEs would differ only in their rank order, and that the mean and SD would (apart from sampling error) be the same. In fact, evidence both old (Lorge & Kruglov, 1953) and new (Humphry et al., 2014) suggests that expert judges tend to think that easy items are harder than they are, and that hard items are easier than they are. That is, the implied scale unit of estimated difficulty tends to be larger (i.e., less discriminating) than the scale unit of empirical difficulty: the judges' estimates are less spread out than the empirical values. Humphry et al. (ibid.) suggested applying a linear transformation to align the scale units, on the assumption that judges are unbiased when estimating passing proportions/probabilities of 50%. Although this assumption seems reasonably plausible, it nevertheless needs empirical support. In any event, we were not confident that we could choose realistic values for scale shrinkage effects to include in our simulation because they may depend on a number of contextual factors. This is an area for further research.

In our simulations, sampling error was the dominant contributor to RMSE, which suggests that attempting to reduce sampling error at the risk of increasing bias may also be worth considering. One way of achieving this would be to apply the 'synthetic linking' approach of Kim et al. (2008) where the final equated cut-score on Test Y is a weighted average of the Test X cut-score and the cut-score derived from the equating. This approach is clearly most suitable when there is some reason to believe that the two tests should have similar cut-scores – perhaps if they have been constructed to the same detailed specification.

The main issue is whether the aggregate of judges' estimates of item difficulty provides useful information about relative test difficulty. The article by Benton (2020, this issue) gives some cause for pessimism here, at least as far as the kind of data we see at GCSE and A Level is concerned. The degree of correlation between judged and empirical item difficulty is clearly an important factor in the usefulness of Angoff-related standard setting methods. Using a small-sample equating method may be preferable to using a standard setting method if typical levels of correlation are to be expected, and indeed this was the conclusion of Dwyer (2016), although it should be noted that the (actual, not simulated) correlations of the judge estimates in his study were in the range 0.39 to 0.49 – lower than observed in many other studies. If it were possible to increase the correlation beyond 0.6 by increasing the number of judges in a judging panel and/or training them to make the mean estimation judgements, then substantial improvements in the accuracy of the standard setting method could be obtained – in the simulation here a correlation of 0.9 was more accurate than the best small-sample equating scenario (a simple random sample of 90 examinees). However, Benton (2020, this issue) argues that rather than focusing on the absolute size of the correlation coefficient, the critical issue is the proportional reduction in error in predicting empirical difficulty from judged difficulty. This takes account of any overall biases and scale differences in judgements as well as disagreements in rank order.

In conclusion, it can be observed that in some contexts standard setting methods are used to achieve the same goal as test equating methods, namely determining cut-scores on test forms that relate to the same performance standard. IRT true-score equating provides a conceptual link between the two, if it is reasonable to conceive of the IRT latent trait as being the same as the abstract continuum containing the performance standard. The simulations reported here have suggested that the overall accuracy of Angoff-based standard setting methods could in some circumstances be similar to what might be expected from test equating with a NEAT design using small samples (N~100) of examinees. Of course, these findings all derive from simulations based on just one dataset, so we are not in a position to make general recommendations about what to do in particular applied contexts. We made choices about how to define the 'true' equating function and which particular standard setting method and small-sample equating method to use, all of which could be varied. The effect of using polytomous items rather than dichotomous anchor items could be explored, as could the effect of varying test length. Furthermore, our method of artificially constructing Tests X and Y ensured that they would be reasonably similar in difficulty. However, these findings point to a way in which practitioners could set up experiments or simulations that more closely match their own particular contexts, in order to discover whether using a standard setting method based on expert judgement might be more accurate than using a small-sample test equating method (or vice

versa); or whether focusing effort on constructing parallel (equally difficult) tests would be a better use of available resource.

References

Benton, T. (2020). How useful is comparative judgement of item difficulty for standard maintaining? *Research Matters: A Cambridge Assessment Publication, 29*, 27–35.

Bramley, T. & Benton, T. (2017). *Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests*. Paper presented at the annual conference of the Association for Educational Assessment-Europe (AEA-Europe), Prague, Czech Republic, 9–11 November, 2017.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*(1), 59–88.

Brennan, R. L. (Ed.). (2006). *Educational Measurement* (4th ed.). Washington, DC: American Council on Education/Praeger.

Cizek, G. J. (1993). Reconsidering Standards and Criteria. *Journal of Educational Measurement, 30*(2), 93–106.

Cizek, G. J., & Earnest, D. S. (2015). Setting performance standards on tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp.212–237). New York: Routledge.

Dwyer, A. C. (2016). Maintaining equivalent cut scores for small sample test forms. *Journal of Educational Measurement, 53*(1), 3–22.

Humphry, S., Heldsinger, S., & Andrich, D. (2014). Requiring a consistent unit of scale between the responses of students and judges in standard setting. *Applied Measurement in Education, 27*(1), 1–18.

Kim, S., von Davier, A. A., & Haberman, S. (2008). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement, 45*(4), 325–342.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices*. (2nd ed.). New York: Springer.

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69–81.

Livingston, S.A., & Kim, S, (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46*, 330–343.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives* (pp.175–217). Mahwah, NJ: Lawrence Erlbaum Associates.

Lorge, I., & Kruglov, L. (1953). The improvement of estimates of test difficulty. *Educational and Psychological Measurement, 13*, 34–46.

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement, 47*(1), 54–75.

Thorndike, R. L. (1982). Item and score conversion by pooled judgment. In P. W. Holland & D. B. Rubin (Eds.), *Test Equating* (pp. 309–317). New York: Academic Press.

How useful is comparative judgement of item difficulty for standard maintaining?

Tom Benton Research Division

Introduction

Developing a way to accurately estimate the relative difficulty of two tests before any students have taken them has long been a holy grail in test development. At one time or another, various organisations have explored how well we can discern the relative difficulties of assessments without actually trialling them with students. Recent research on this topic has been produced by Cito in the Netherlands (van Onna, Lampe, & Crompvoets, 2019), ETS in the United States (Attali, Saldivia, Jackson, Schuppan, & Wanamaker, 2014) and Cambridge Assessment in the UK (Curcin, Black, & Bramley, 2009). Item trialling is often undesirable as it places some of the burden of test development upon schools and students, and can lead to concerns over the security of items.

If accurate predictions of item difficulty were possible then, in the context of UK examinations, this would mean being able to accurately set grade boundaries for this year's GCSE exams before any students have attempted the paper. It would also provide an alternative to the current approach of "comparable outcomes" to awarding and its inherent implication that (broadly speaking) the percentage of pupils achieving high grades will not change from the previous year (Benton, 2016). Outside of the UK context, being able to accurately predict the difficulty of items might allow "lowering the sample sizes required for item pretesting, leading to lower costs and increased security of items" (Attali et al., 2014, p.7).

The previous article (Bramley, 2020) has considered the extent to which a particular form of expert judgement (the 'mean estimation' variant of the Angoff Method) might provide sufficiently accurate information on the relative difficulty of two tests. The present article explores the value of expert judgements of item difficulties derived in a different manner – by comparative judgement (CJ).

In this context, a CJ study requires expert judges to sort sets of items according to their perceived difficulty (PD). The rationale for using CJ is that previous research has indicated that judges tend to "be good at predicting the relative difficulties of items but not absolute levels" (Mislevy, Sheehan, & Wingersky, 1993, p.59). Placing items in a rank order of difficulty is conceivably a more intuitive task then estimating the proportion of minimally competent candidates who will answer them correctly, as must be done under the Angoff Method. As such,

a CJ approach may be considered likely to provide better estimates of the relative difficulty of items (Attali et al., 2014).

One type of CJ exercise is a pairwise comparison study where judges are shown two items at a time, and must simply decide which of the pair is more difficult to answer correctly (see, for example, Ofqual, 2015). An alternative approach is rank ordering. As an example of this method, Curcin et al. (2009) presented judges with packs of four items which they had to place into order of difficulty. In either a pairwise comparison or a rank ordering study, each item is typically included in multiple different comparisons undertaken by different judges. The results from all the judgements of all the items by all the judges are combined into a single data set and analysed using the Bradley-Terry model (or similar) to place all of the items on a continuous scale from the easiest to the most difficult. The position of each item on this scale gives a value of PD that might then be used to allow us to infer the relative difficulty of two tests overall.

Of course, it is possible to use a rank ordering approach to judging item difficulty without the need to employ the Bradley-Terry model. For example, perhaps the earliest rank ordering study of this kind (Lorge & Kruglov, 1952) required judges to review all of the items across two tests at once and place them into a single rank order. Having done this, the average rank assigned to a given item across all of the judges provides an estimate of PD.

Existing literature suggests several ways in which estimates of PD from a CJ study might be used to infer the relative difficulty of tests. Usually these rely on linking the estimates of PD for each item to empirical difficulties as defined using item response theory (IRT) or Rasch analysis. An extreme example of this approach, taken by Holmes, Meadows and Stockford (2018) and Bramley (2010), is simply to use estimates of PD from a Bradley-Terry model directly as substitutes for empirical estimates of Rasch difficulty. There seems little justification for this approach. The former relate to the probability of a judge considering one item in a pair more difficult than another, whereas the latter relate to the probability of students answering an item correctly. These are clearly distinct concepts and there is no obvious justification for using one as a substitute for the other. In other cases, PD is used as an input to a statistical model to help infer the likely location of empirical item difficulties (e.g., Mislevy et al., 1993) or to calibrate item difficulties that have been separately estimated for two tests onto the same scale (van Onna et al., 2019).

Two immediate problems occur in attempting to relate estimates of PD to IRT difficulty parameters. Firstly, from a practical point of view, it is not obvious how items with more than one mark available should be treated. CJ studies will usually only provide a single value for the PD of each item but in order to use IRT it is necessary to estimate the difficulty of each mark within the item. Secondly, the use of IRT makes it difficult to either calculate or communicate the likely accuracy with which such methods can actually equate two tests.

The aim of this article is to simplify the evidence on the value of comparative judgements of item difficulty for estimating the overall difficulty of tests. To begin with, I will review the evidence on the strength of the relationships between estimates of item difficulty derived from CJ and actual empirical difficulties. After this, I will show how we can combine perceived item difficulties with simple (non-IRT) statistical methods to estimate the relative difficulty of two tests. Crucially, the simple approach will also allow us to assess exactly how accurate equating tests based purely on PD is likely to be in general.

How strong is the relationship between estimates of PD from paired comparisons and empirical item difficulty?

To investigate the relationship between PD derived from a CJ study and actual item difficulties, I used data from The Office of Qualifications and Examinations Regulation (Ofqual, 2015). This study of the relative difficulty of various Mathematics exams included items from six legacy GCSEs in Mathematics offered by OCR. The estimates of PD of each item were published as part of the study (Ofqual, 2015, Appendix B, pp.140–146) and it was possible to link them to empirical data on the performance of students on the same questions and evaluate the strength of the association. Each of the assessments was taken by more than 5,000 candidates, providing ample data for empirical estimates of item difficulty.

Table 1 provides further details of the assessments included in analysis. They each contained between 20 and 40 items. Each test contained both single-mark (dichotomous) and multi-mark (polytomous) items. The empirical difficulty of each item was estimated using its facility. Item facilities are usually presented on a scale from 0 to 100, and represent the mean score on an item expressed as a percentage of the maximum score available. For the items in these six tests the mean facility was close to 50% meaning that, for a typical item, candidates achieved about half of the available marks on average. The standard deviations (SD) of the facilities across items are also shown in Table 1. These show that, although the average facility was generally close to 50% in each test, items with a wide range of difficulties were included.

The most common way to evaluate the strength of the association

Table 1: Correlations between PD and facility for the six Mathematics GCSE papers

Unit	*Tier*	*Number of items*	*Number of marks*	*Mean Facility*	*SD of Facilities*	*Correlation of PD and Facility*	*Residual SD of Facilities*
Unit 1	Foundation	30	60	47.9	28.2	-0.57	23.6
Unit 1	Higher	27	60	44.9	19.8	-0.44	18.1
Unit 2	Foundation	28	60	45.6	21.1	-0.50	18.6
Unit 2	Higher	22	58[1]	50.6	20.9	-0.26	20.6
Unit 3	Foundation	39	100	58.2	23.6	-0.57	19.7
Unit 3	Higher	35	100	68.7	22.2	-0.48	19.8

1. The original test had 23 items and 60 marks available. However, one item was omitted from Ofqual's study and so only 22 items and 58 marks are included here.

between two quantities is to calculate a correlation coefficient. The (Pearson) correlations between PD and empirical facilities are also shown in Table 1[2]. The negative sign of these correlations is expected – items that are perceived to be more difficult are answered correctly less often.

Of more interest is the size of these correlations. In Table 1 the sizes of the correlations between PD and facility range from 0.26 to 0.57. However, it is not immediately clear how to interpret these values. Clearly, there is some relationship between the perceived and actual difficulty of items. However, is the relationship strong enough to be of any value in judging the relative difficulty of, and ultimately in equating, two tests?

To investigate this, I compared the strength of the correlations in Table 1 to the correlations between the overall item facilities and facilities based on very small samples of candidates. For example, for any of the above tests, we might select just one candidate. Then, for each item, the facility (based on this one candidate) is zero if they get the item completely wrong, 100 if they get it completely right, and something in between if they achieve some but not all of the marks. The correlation between the item facilities based on this one candidate and item facilities based on the full population can then be calculated. The procedure was repeated 100 times[3] for each of the six assessments to get an idea of what correlation between a facility from one candidate and the overall facility we might expect. The same method was then repeated to estimate the predictive value of data from small samples of two, three, four, five, six, seven, eight, nine and ten candidates.

The results are shown in Figure 1. The dotted lines represent the size of the correlations between PD and facility for each test (see Table 1). The boxplots show the distribution of correlations between facilities from small samples and overall facilities for each sample size across the replications. As can be seen, in most cases, the correlation between PD (based on a CJ study) and facility is very similar to what we might expect to achieve on average by using a sample of *just one candidate*. With a sample of five we can virtually guarantee that the data from even so few candidates will be more predictive of actual item difficulty than a

2. Spearman correlations were of very similar magnitude and (for brevity) are not shown.

3. That is, with 100 separate candidates.

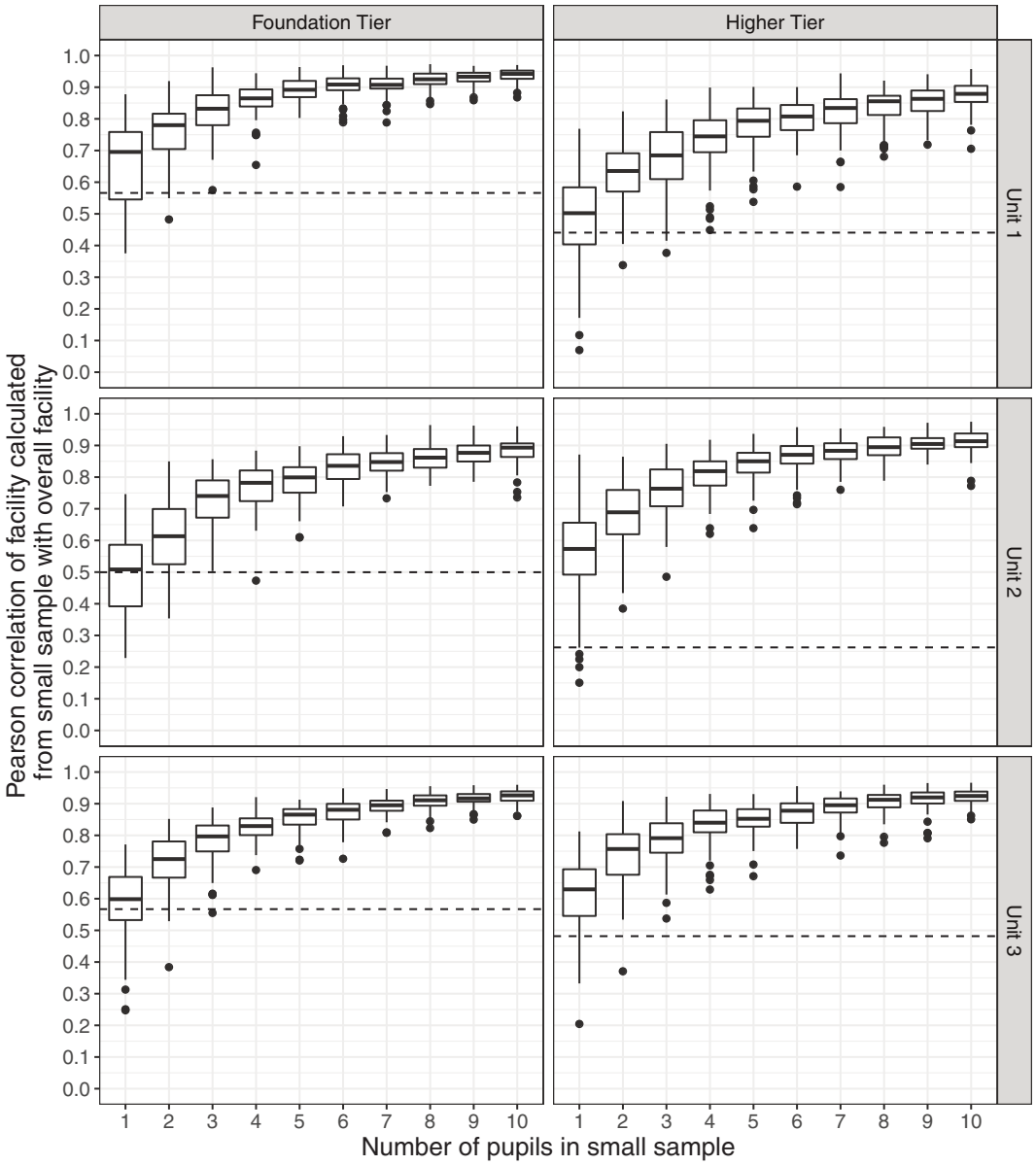


Figure 1: Correlations between facilities calculated from small subsamples and full sample facilities for six GCSE Mathematics assessments. The dotted lines show the size of the Pearson correlation between PD and facility in each case.

CJ study. This immediately suggests that CJ exercises to estimate PD are a very weak source of evidence about the difficulty of items.

An alternative to correlations – the actual accuracy of predictions

The above analysis suggests that PD cannot be seen as a strong form of evidence. However, it does not necessarily mean such information is useless. Were it possible to get even a single typical student with the correct level of exam preparation and motivation to trial a paper before it goes live, and this could be done without any security concerns, this would likely be considered a very useful resource to test developers. However, in reality this is tricky. For this reason, despite the above results, it is still of interest to explore the accuracy with which PD can predict actual item difficulty in more detail.

Like the analysis above, most studies evaluating the predictive value of PD focus upon correlation coefficients. However, looking at correlations alone can provide a misleading picture. The reason for this is that correlations will tend to increase along with the spread of values included in the study. For example, imagine that some basic Mathematics questions asking students to add two single digit numbers had been added to the above exams. Such questions would have been self-evidently easier than any other items in the exams and the vast majority of students would have answered them correctly. Thus, including such questions would make it easier for experts to correctly distinguish the relative empirical difficulty of at least some of the items, and the correlations between PD and facility would increase.

This same effect exists (perhaps in a less extreme form) whenever we use correlations to assess the strength of associations. In the current context, the greater the spread of actual item difficulties within a test, the easier it will be for experts to discern this, and the higher the correlation between PD and facility will be. Ultimately, a more useful way to understand the value of PD is to actually calculate how accurately we can predict item facilities. That is, if we were to use PD to predict the likely facility of a new item (for the same population of students), how close would that prediction be to the actual facility?

Because it is helpful for the calculations that follow in the next section, rather than evaluating the average size of differences between predicted and actual values (the mean absolute differences), we will actually use the square root of the mean squared differences (the residual standard deviation). Residual standard deviations are higher than mean absolute differences but, very broadly speaking, can be interpreted in the same way in that both give an idea of the typical difference between predicted and actual facilities.

Residual standard deviations of facilities given PD are provided in the final column of Table 1. They can actually be calculated from the overall standard deviation of facilities in each test, the correlation between facility and PD, and the number of items in the test using the formula below:

$$Residual\ SD\ of\ Facility = SD(Facility)\ \sqrt{\frac{(1-r^2)(n-1)}{(n-2)}} \qquad (1)$$

where r is the correlation between PD and facility, and n is the number of items in the test. The terms relating to the number of items in the above formula adjust for the fact that a regression line will be fitted to the existing data on actual facilities. As such, without this correction, we would probably overestimate the likely accuracy of future predictions in brand new data sets.

The values in Table 1 indicate that PD allows empirical facilities to be predicted to within a little under 20% on average[4]. Crucially, by comparing these values to the overall standard deviation of facilities in each test (as Table 1 shows, these are around 22), we can see that this level of accuracy is only marginally better than could be achieved by simply guessing that all items would have a facility near the average for that test. That is, having a value for the PD of each item only marginally improves the accuracy with which we can predict empirical difficulty. We can also see that the tests displaying the highest correlation between PD and facility are not associated with predictions of facility actually being more accurate. For example, the Unit 1 Foundation tier test displayed one of the highest correlations between PD and facility but also had the highest residual standard deviations (i.e., the worst predictive accuracy).

Figure 2 allows a visual exploration of the same idea. The charts show the associations between PD and facility with a regression line shown in blue in each case. We can see that, within each assessment there is a clear relationship between PD and facility. However, there is nearly the same spread in empirical facilities for any fixed value of PD as there is overall. Around a third of the empirical facilities in Figure 2 are more than 20 percentage points away from what would be predicted based on PD.

The above results illustrate the problem of relying on correlations alone to assess the value of PD. To illustrate this further, analysis was also undertaken evaluating the association between PD of items and the percentage of candidates that answered each item fully correctly (i.e., achieved all of the available marks). This analysis showed that the correlations increased and now ranged between -0.39 and -0.73. This could be taken as indicating that PD was more predictive of the percentage of candidates answering items fully correctly than of facility. However, further analysis revealed that the actual accuracies of predictions were, in fact, worse than the accuracies of predictions of item facility (shown in Table 1). Specifically, even as the correlations increased, the residual standard deviations also increased. In other words, reporting correlations alone may not give the most appropriate picture of the value of PD.

The actual accuracy of predictions based on PD for previous studies

The results above give a potentially disappointing picture of the accuracy with which PD can predict empirical difficulty. This is in contrast to some academic literature on this subject which presents a more positive picture of the potential of PD. To investigate this discrepancy further, results from a number of previous studies of this type are shown in Table 2. These results should not be taken as representing a systematic review of all of the articles on this issue. They simply reflect a number of articles that I am familiar with, presenting a range of views on the value of PD.

All of the studies listed in Table 2 except one (Humphry, Heldsinger, & Andrich, 2014) made use of a CJ approach to eliciting item difficulties. This exception was included simply to represent the fact that studies using methods other than CJ also occasionally claim that judges can successfully estimate the relative difficulty of items and to ensure that at least one such non-CJ study was subjected to some scrutiny.

For each study in Table 2, I have identified the reported correlation between PD and facility. In several cases, the authors reported Spearman

4. The mean absolute difference between predicted and actual facilities is 16%..

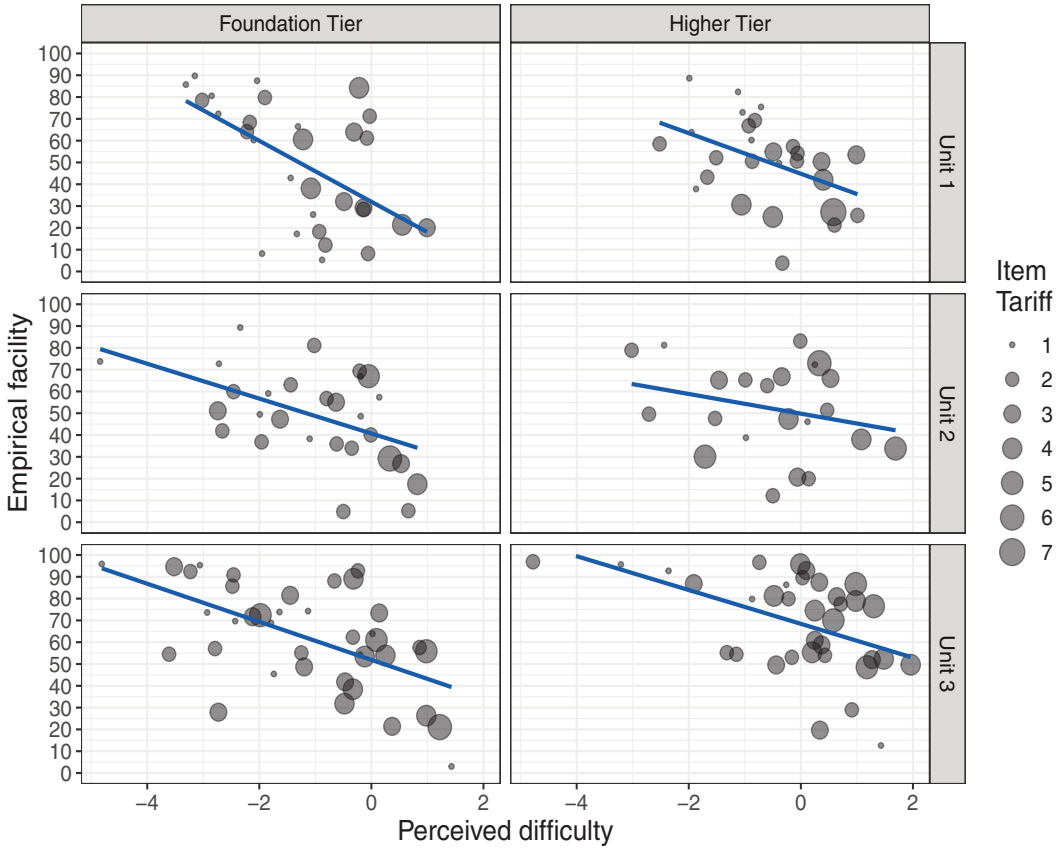


Figure 2: Relationship between PD and item facility for each Mathematics GCSE unit.

(i.e., rank correlations) rather than Pearson correlations, however, this should not make a major difference to results. Indeed, in several cases, by analysing data within scatterplots I was able to verify that the Pearson correlations would give similar values. The article by Humphry et al. (2014) did not present the correlation at all and it was necessary to estimate its value based on data contained within a figure in the article. As can be seen, the values of the correlations vary considerably between studies. For example, on the one hand, the study by Attali et al. (2014) found a correlation between PD and empirical difficulty close to 0.8, leading the authors to conclude that "contrary to previous investigations, judges are able to discriminate quite well between easier and harder items when they are given a comparative judgment task" (p.6). At the other end of the spectrum, Curcin et al. (2009) found a correlation of just 0.18 between PD and facility for a particular multiple choice test and were left to "speculate why rank ordering failed to elicit consistently valid judgements about question difficulty" (p.6). Thus, focussing purely on correlations gives the impression of CJ sometimes being a highly effective means to estimate item difficulty and sometimes being almost entirely ineffective.

Table 2 also provides the standard deviations of the facilities of items included in each study and, based on these and the formula provided above, the estimated residual standard deviation of facilities. Note that

Table 2: Correlations and residual SD of facilities for studies in existing literature

Author	*Assessments studied*	*Method*	*Number of items*	*SD of Facilities*	*Correlation of PD and Facility*	*Residual SD of Facilities*
Humphry et al. (2014)	Multiple choice Year 7 reading test	Angoff procedure	35	25.3	-0.80	15.4
Lorge and Kruglov (1952)	High school admissions tests in arithmetic	All items ranked by four judges. Average the item rankings.	86	25.1	-0.84	13.7
Attali et al. (2014)	SAT Mathematics	Ranking of seven items by one judge	7	28.2	-0.79[5]	18.9
Ofqual (2018)	AS Level Mathematics	Formal pairwise comparisons	554	19.2	-0.49	16.8
Ofqual (2017)	GCSE Biology	Formal pairwise comparisons	351	23.4	-0.50	20.3
Curcin et al. (2009)	Multiple choice test in Administration	Formal rank ordering	30	22.4	-0.34	21.4
Curcin et al. (2009)	Multiple choice test in Road Haulage and Passenger Transport	Formal rank ordering	30	16.9	-0.18	16.9

5. Median value for correlation across 825 packs of seven items each of which were assessed by a single judge.

in only one of the studies (Lorge & Kruglov, 1952) was the standard deviation of item facilities reported in the original work. In the other cases, it was necessary to either calculate the standard deviation by reading the empirical data for individual items from plots, or to estimate it by careful reading of the information that was provided. For this reason, although Curcin et al. (2009) studied eight separate tests, only two indicative examples are included in Table 2. Similarly, although Ofqual (2017) analysed six separate Science assessments (two Biology, two Chemistry and two Physics tests), with correlations between PD and facility ranging from -0.50 to -0.36 (see Ofqual, 2017, Appendix C), it was only possible to include a single Biology test in the analysis here.

Studying the residual standard deviation of facilities given PD leads to a very different set of conclusions to looking at correlations alone. In particular, we can see that, despite the range of correlations in Table 2, there is far less difference in the actual accuracy with which PD could predict facility in each study. In particular, we can see that the high correlation reported by Attali et al. (2014) was associated with a residual standard deviation of 18.9, whereas for the study with the low correlation reported by Curcin et al. (2009) the residual standard deviation was 16.9. That is, contrary to what we might expect given the tone of the conclusions, the latter study was able to produce more accurate predictions of item facilities than the former.

The point here is not that PD is always equally predictive of actual difficulty. It is perfectly plausible that it is easier for judges to assess the relative difficulty of arithmetic questions aimed at pre-high school children (Lorge & Kruglov, 1952) than it is to perform the same task for Biology questions aimed at older teenagers (Ofqual, 2017). However, it is clear that previous attempts to use PD cannot be classified as successes or failures based on the correlation between PD and facility alone. More importantly, the actual accuracy with which PD can predict empirical difficulty is similar enough across both previous studies (Table 2) and the data sets we are analysing in this article (Table 1), that the following sections regarding the accuracy with which we can equate tests based on PD should generalise reasonably well across contexts beyond GCSE Mathematics.

How accurately can we equate tests using PD?

We have seen above that PD is not a particularly good predictor of empirical difficulty and that the level of accuracy is fairly consistent across different studies. However, it might be hoped that, once PDs are aggregated across items to whole tests, the various errors will cancel out, leading to a good way of judging the relative difficulty of assessments. For example, Holmes et al. (2018) conclude that "providing there are no systematic biases in judging expected difficulty of items from different exam boards, the median and spread of predicted item difficulty for a paper will represent the actual difficulty of that paper reasonably well" (p.386).

This section considers this question in more detail. That is, given the accuracy with which we can predict item facilities, how well can we predict the difficulty of a whole test? For simplicity, we will imagine that the difficulty of a test is adequately represented by the mean score that would be achieved on it by a given population of students. For example, imagine we knew that the mean score on last year's test was 50 out of 100. Then, using the PD of items and a predictive model devised using the previous year's empirical data, we predicted that for the same set of students, the mean score of this year's test would be 55 out of 100. We might reasonably conclude that this year's test was five marks easier than last year's. Thus, we might expect that grade boundaries should be about five marks higher this year than last year. Such reasoning is the basis for a type of statistical equating (i.e., calibrating tests against one another), known as mean equating. For grade boundaries that are reasonably close to the mean this approach is fairly easy to justify. If grade boundaries are a long way from the mean then this method is less justifiable. However, it may still provide a useful starting source of evidence. For example, it might be used as an input to a more sophisticated small-sample approach, such as circle-arc equating (Livingston & Kim, 2009). Either way, exploring the accuracy with which we expect to be able to predict the mean test scores for a fixed population provides a simple mechanism to help us explore the value of PD for equating tests.

On the basis of the above argument, we approximate how accurately we can equate tests using PD by the accuracy with which we can predict test means. Specifically, we want to calculate the standard error of a predicted test mean based upon the PDs of all the items in the test – that is, the expected root mean square error. To begin with, we note that the mean score on a whole test is simply the sum of the mean scores on the individual items. That is,

$$Predicted\ test\ mean = \sum Predicted\ item\ mean \qquad (2)$$

Next, if we assume a best case scenario that errors in predicted item means are independent (as opposed to consistently systematically biased), then the squared error of the predicted test mean (technically referred to as the "variance") will be equal to the sum of the squared error in the predicted item means. Mathematically, we write these concepts as:

$$SE(predicted\ test\ mean) = \sqrt{Variance(test\ mean|PDs)} = \sqrt{\sum Variance(item\ mean|PD)} \qquad (3)$$

Next, we note that the mean score on an item is just its facility (divided by 100) multiplied by the number of available marks. As such, the squared error in the item mean will be the square of the error in the facility multiplied by the maximum number of marks. From Tables 1 and 2 we can see that the residual standard deviation of item facilities given PDs tends to be about 20 (20% of the item maxima). Thus, the expected squared error of the mean score on an item given PD will be equal to $(0.2 * item\ max)^2$. Thus, continuing the mathematical formula above, the standard error with which we can predict the mean score on a test using item PDs is approximated by:

$$SE(predicted\ test\ mean) \approx \sqrt{\sum (0.2 * item\ max)^2} = 0.2\sqrt{\sum item\ max^2} = \frac{RSSIM}{5} \qquad (4)$$

The final term in the above equating (the RSSIM) was introduced in Benton (2019) and is just the square root of the sum of the squared item maxima. Its occurrence in the above formula relates to the fact that we would expect to be able to predict mean scores from item PDs more accurately for a test consisting of many low tariff items than that for one consisting of a few items worth many marks. This makes sense as if we have a greater number of items there is more chance for errors in predicted item means to cancel out.

The above formula suggests that the accuracy with which we can

predict test means based on item PDs is approximated by the RSSIM divided by five. Table 3 illustrates the implications of this formula for the six GCSE Mathematics tests in analysis. For each of the tests we have calculated the RSSIM. For example, the Unit 1 Foundation tier test consisted of twelve 1-mark items, ten 2-mark items, four 3-mark items and four 4-mark items. Thus, the RSSIM was equal to:

$$\sqrt{(12 * 1^2 + 10 * 2^2 + 4 * 3^2 + 4 * 4^2} = \sqrt{152} = 12.3.$$

Using the above formula, we can estimate the expected standard error of a predicted test mean as a fifth of this value. For example, for the Foundation tier Unit 1, we estimate that the standard error of predicted test mean is 12.3/5=2.5 marks. For this same test, using common statistical practice, we can infer that a 95% confidence interval for the predicted mean would cover a range of plus or minus double this standard error, that is plus or minus roughly five marks. Thus, by assuming that the accuracy with which we can predict the mean gives a good approximation to how accurately we can position grade boundaries; we infer that on the basis of PD alone, the grade boundaries for this test might reasonably be set anywhere within a range of ten marks. Given that this whole test has a maximum of just 60 marks, this is only the vaguest sense of where grade boundaries should be placed.

Table 3: Expected accuracies with PD

Unit	*Tier*	*Number of items*	*Number of marks*	*RSSIM*	*Standard errors on predicted boundaries*	*Width of 95% confidence interval for boundaries*
Unit 1	Foundation	30	60	12.3	2.5	9.7
Unit 1	Higher	27	60	13.4	2.7	10.5
Unit 2	Foundation	28	60	13.0	2.6	10.2
Unit 2	Higher	22	58	13.9	2.8	10.9
Unit 3	Foundation	39	100	18.2	3.6	14.2
Unit 3	Higher	35	100	18.3	3.7	14.4

Further illustration using split halves

This section provides an empirical illustration of the extent to which PDs of items might allow an accurate assessment of the relative difficulty of tests. To explore this, I used data from the Foundation tier Unit 3 paper. This paper was chosen as it included the largest number of items and also displayed the largest correlation between perceived and empirical item difficulties (see Table 1).

The items within this paper were randomly split into two half-papers, such that each half-paper consisted of 50 marks. The empirical mean score of each half-test was calculated to indicate the actual relative difficulty of the two tests. The weighted mean PD was also calculated for each test with more weight given to items worth a larger number of marks[6]. Giving more weight to PDs of items with more marks reflects the way we would use PD to predict mean test scores. The weighted mean PD provided an indication of the overall PD of each half-test. This process (split items into half-tests; calculate mean scores; calculate mean weighted PDs) was then repeated 100 times.

Across the 100 replications, Figure 3 compares the difference in PD of half-tests to the difference in means (i.e., actual difficulty). As can be seen, there is clearly some relationship between perceived and actual difficulty but it could hardly be described as providing an accurate basis for equating. For example, it is clear that where half-tests are of equal PD, there are instances of one half-test being around five marks

6. Using the weighted median PD was also trialled but was found to make no substantial difference to the final results.

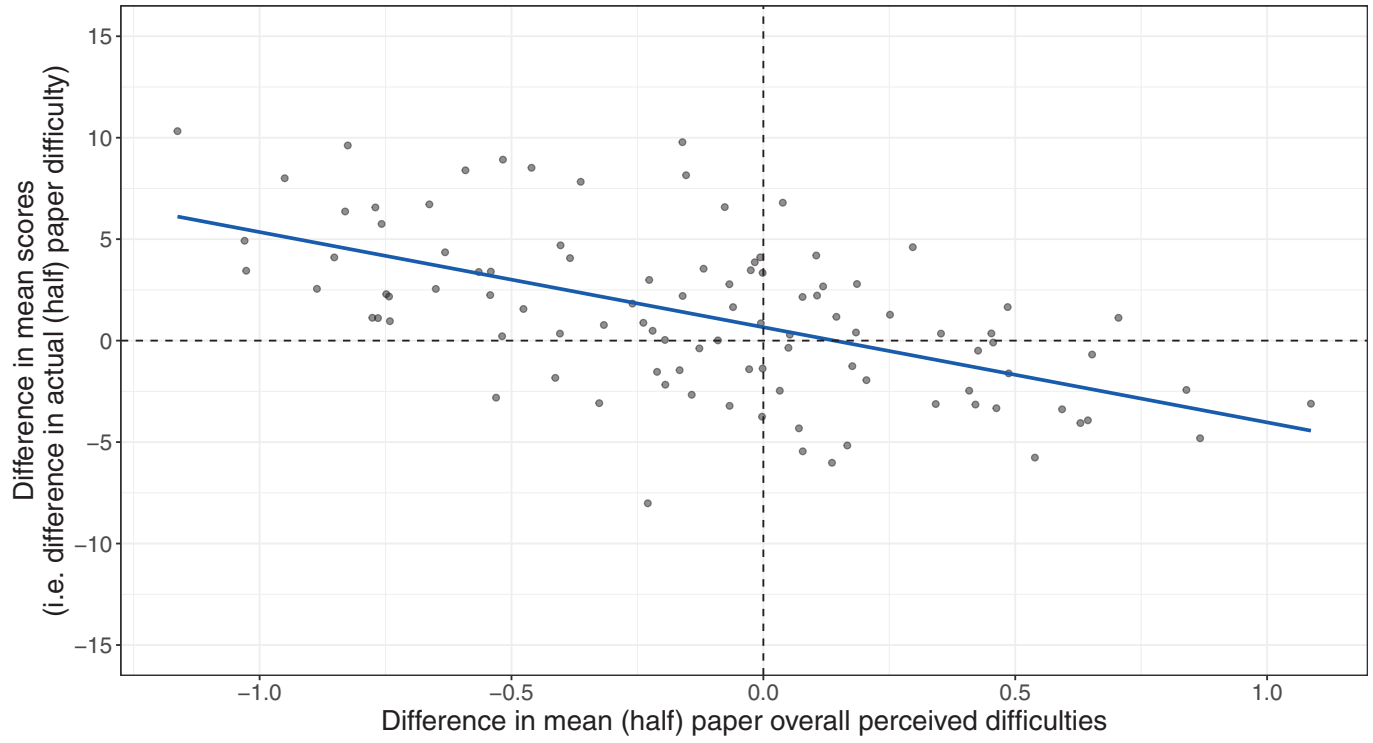


Figure 3: The relationship between differences in mean PD of two half-papers and differences in empirical mean scores. A fitted regression line is shown in blue.

(out of 50) easier than the other. Furthermore, PD does not even provide a reliable idea of the direction of difference in difficulty. Specifically, in 28 out of the 100 cases analysed, the direction of difference in PD was inconsistent with the direction of difference in means (i.e., a half-paper was perceived as more difficult when it was actually easier).

Conclusion

This article has shown that PD is not a particularly good predictor of actual item difficulties. Specifically:

- On average, the predictive value of PD derived from a CJ exercise is roughly equivalent to the value of empirical data from just one student.
- For a fixed level of PD, item facilities have a standard deviation of about 20 percentage points. In other words, items that are perceived as equally difficult can have substantially different empirical difficulties.
- Broadly speaking, this level of predictive accuracy holds even for existing studies that have reported high correlations between perceived and empirical difficulty.

The analysis comparing the relative value of small samples of students and expert judgement suggests the intriguing possibility that we may generate better evidence if the subject experts involved in studies of item difficulty were replaced with the same number of students taking part in item trials. It may be argued that, because students would be far less well prepared and motivated in such an exercise than in a high-stakes exam, this would not provide an accurate idea of the actual relative difficulty of items. However, this view is not necessarily supported by the evidence. For example, in developing tests for use in primary school schools in England, the Standards and Testing Agency (STA) routinely trials all items with around one thousand pupils before they are used in live settings (STA, 2018). Estimates of item difficulty from these low-stakes trials are very highly correlated with difficulties estimated using live exam data. For example, for the Key Stage 2 Mathematics test taken in 2017 the correlation between IRT difficulty thresholds estimated from the two different sources of data was 0.976[7].

This article has also shown that it is not correct to assume that, when aggregated to the level of whole test papers, PD will provide an accurate means to judge the relative difficulty of two assessments. Equation 4 shows how to estimate the likely reliability of setting grade boundaries based on PD. For the GCSE Mathematics tests explored in this article the formula suggested that, given item PDs, a given grade boundary could reasonably be positioned at any score across a range covering roughly one sixth of the maximum available. This estimated level of accuracy is consistent with the simulations reported in Bramley (2020) for the situation where Angoff judgements of difficulty and actual item difficulties have a correlation of 0.6. This level of precision cannot be described as an accurate idea of where grade boundaries should be positioned. It should be noted that even achieving this level of accuracy requires an assumption that the items in the tests being compared come from the same "population" in some sense, and that there is no systematic reason for the items in one test being harder (relative to PD) than those in another. In practice, such systematic biases can occur. For example, Ofqual (2017, Appendix A) provides an example where, given equal PD, items produced by one assessment agency were systematically harder than those produced by another. As such, the formula provided in this article should be seen as giving a best-case view of the accuracy of the approach.

Finally, it is worth noting that, within each of the GCSE Mathematics tests that were analysed, variations in item facilities were barely any lower for fixed levels of PD than they were overall. That is, simply guessing that each item would display an average level of difficulty for the given test provides nearly as accurate a prediction of individual item difficulties as a full CJ exercise to elicit PDs. This, in turn, implies that knowing the PDs of new items, and the relationship between PD and empirical difficulty in the past, is hardly any more informative than just knowing the average difficulty of items within a particular paper historically.

Thinking about this from a practical perspective, the issue we are trying to solve is that we do not know how difficult the questions in current exam papers are. However, we do know how difficult they tended to be in the past. The results in this article indicate that PD does not add much new useful information to this. Therefore, we must conclude that the accuracy of using PDs to set grade boundaries is hardly any different from simply assuming that tests made to the same design specifications will always be equally difficult over time and, thus, that grade boundaries should remained fixed. The idea of using fixed grade boundaries over time has been suggested before (Bramley, 2012; Bramley, 2018). Whilst I am not necessarily recommending such an approach, it has to be conceded that it is hard to be in favour of the use of PDs for setting grade boundaries whilst objecting to the use of fixed grade boundaries.

In fact, for GCSEs and A Levels in England, PD already plays some role in the creation of test papers. Specifically, item writers are already required to identify the target level of each item. Usually this is expressed in terms of the grades available on the test and indicates the expected level of skill required to answer the question. As items are assembled into tests, a specification grid is used to ensure that the proportion of items at each level (as well as the balance of different topics) is kept consistent from year to year. Thus, a mechanism already exists to ensure that PD should remain reasonably constant over time, and, as such, we might expect the grade boundaries to remain constant. Given this, setting grade boundaries using CJ of the perceived difficulties of items could be seen as an expensive way of deciding that we ought to keep grade boundaries fixed over time. Whether this is a good idea is a wider question for further research.

7. Correlation provided by the Standards and Testing Agency under the Freedom of Information Act 2000.

References

Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating Item Difficulty With Comparative Judgments. *ETS Research Report No. RR-14-39*. Princeton, NJ: Educational Testing Service. Retrieved from: http://dx.doi.org/10.1002/ets2.12042.

Benton, T. (2016). Comparable Outcomes: Scourge or Scapegoat? Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Benton, T. (2019). Which is better: one experienced marker or many inexperienced markers? *Research Matters: A Cambridge Assessment publication, 28*, 2–10.

Bramley, T. (2010). *Locating objects on a latent trait using Rasch analysis of experts' judgments*. Paper presented at the conference "Probabilistic models for measurement in education, psychology, social science and health", Copenhagen, Denmark, June 2010.

Bramley, T. (2012). *What if the grade boundaries on all A level examinations were set at a fixed proportion of the total mark?* Paper presented at the Maintaining Examination Standards seminar, London.

Bramley, T. (2018). When can a case be made for using fixed pass marks? *Research Matters: A Cambridge Assessment publication, 25*, 8–13.

Bramley, T. (2020). Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests. *Research Matters: A Cambridge Assessment publication, 29*, 23–27.

Curcin, M., Black, B. & Bramley, T. (2009). *Standard maintaining by expert judgment on multiple-choice tests: a new use for the rank-ordering method*. Paper presented at the British Educational Research Association Annual Conference, University of Manchester, 2–5 September 2009.

Holmes, S. D., Meadows, M., Stockford, I., & He, Q. (2018). Investigating the Comparability of Examination Difficulty Using Comparative Judgement and Rasch Modelling. *International Journal of Testing, 18*(4), 366–391.

Humphry, S., Heldsinger, S., & Andrich, D. (2014). Requiring a Consistent Unit of Scale Between the Responses of Students and Judges in Standard Setting, *Applied Measurement in Education, 27*(1), 1–18.

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46*(3), 330–343.

Lorge, I., & Kruglov, L. (1952). A suggested technique for the improvement of difficulty prediction of test items. *Educational and Psychological Measurement, 12*(4), 554–561.

Mislevy, R.J., Sheehan, K.M., & Wingersky, M. (1993). How to equate tests with little or no data, *Journal of Educational Measurement, 30*(1), 55–78.

Ofqual (2015). *A comparison of expected difficulty, actual difficulty, and assessment of problem solving across GCSE Maths sample assessment materials*. Ofqual/15/5679. Coventry: Ofqual. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/429117/2015-05-21-gcse-maths-research-on-sample-assessment-materials.pdf.

Ofqual (2017). *GCSE science: An evaluation of the expected difficulty of items*. Ofqual/17/6163. Coventry: Ofqual. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/592709/gcse-science-an-evaluation-of-the-expected-difficulty-of-items.pdf.

Ofqual (2018). *A level and AS mathematics: An evaluation of the expected item difficulty*. Ofqual/18/6344. Coventry: Ofqual. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/676730/A_level_and_AS_mathematics_An_evaluation_of_the_expected_item_difficulty.pdf.

STA (2018). National curriculum test handbook: 2018 Key stages 1 and 2. STA/19/8312/e. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/765749/2018_NCT_Handbook_PDFA.pdf.

van Onna, M., Lampe, T., & Crompvoets, E. (2019). *Equating by pairwise comparison*. Presentation at the 20th annual AEA-Europe conference, Lisbon, Portugal.

Research News

Anouk Peigne Research Division

Conference Presentations

British Educational Research Association (BERA)

The annual conference of the British Educational Research Association took place in Manchester in September 2019 and allowed researchers to discuss findings across many educational themes:

David Beauchamp and Filio Constantinou (Research Division): *To what extent is the language of this test question readable? Tools for investigating the linguistic accessibility of assessment material.*

Matthew Carroll (Research Division): *Longitudinal data in education research.*

Vicki Crisp (Research Division): *Context in science exams.*

Nicky Rushton (Research Division): *Teachers' use of and views about enquiry-based learning in the new 9-1 GCSE Geography specifications.*

Carmen Vidal Rodeiro (Research Division): *How does A-level subject choice and students background characteristics relate to Higher Education participation?*

Emma Walland: *Teacher decision-making on post-16 provision in response to reform.*

British Psychological Society East of England

This conference took place at the University of Anglia Ruskin, Cambridge, UK, in September. Professionals gathered in a mixture of workshops, oral and poster presentations to discuss around this year's theme: *The Psychology of Wellbeing*. The following paper was presented:

Irenka Suto (Research Division): *It's Time to Talk about talking about research; Presentation anxiety and other aspects of our jobs which make researchers tense.* This was based on research with her colleague Gill Elliott.

International Society for Design and Development in Education

The 15th annual conference of the International Society for Design and Development in Education took place in Pittsburgh, USA. The theme was *Design for the Future* and attendees participated in group work sessions, presentations, talks and informal conversations. The following paper was presented:

Martin Johnson (Research Division): *Development Challenges in Challenging Contexts: A story of EiE curriculum framework development*. This was based on his research with colleagues Tori Coleman and Sinéad Fitzsimons.

Association for Educational Assessment-Europe (AEA-Europe)

The AEA-Europe annual conference took place in Lisbon, Portugal, in November 2019. The conference's topic was *Assessment for transformation: teaching, learning and improving educational outcomes.* Various researchers from Cambridge Assessment presented papers:

Tom Bramley, co-authored with Victoria Crisp (Research Division): *Spoilt for choice? Is it a good idea to let students choose which questions they answer in an exam?*

Gill Elliot, co-authored with Jo Ireland (Research Division): *Re-heated meals: Revisiting the teaching, learning and assessment of practical cookery in schools.*

Filio Constantinou (Research Division): *Tests as texts: investigative text questions from a sociolinguistic perspective.*

Filio Constantinou, co-authored with David Beauchamp (Research Division): *To what extent is the language of this test question readable? Tools for investigating the linguistic accessibility of assessment material.*

Martina Kuvalja (Cambridge English), Stuart Shaw (Cambridge Assessment International Education), co-authored with Sarah Mattey (Research Division) and Giota Petkaki (Cambridge Assessment International Education): *Assessment of problem-solving skills.*

Isabel Nisbet and Stuart Shaw (Cambridge Assessment International Education): workshop *Is assessment fair?*

Martin Johnson, co-authored with Victoria Coleman (Research Division): *Getting out of their heads – using concept maps to elicit teachers' assessment literacy.*

Carla Pastorino (Cambridge Assessment International Education): *Student engagement with on-screen assessments: A systematic literature review.*

Alison Rodrigues (Cambridge Assessment International Education) and Sarah Hugues (OCR): *From opinion to evidence: transforming organisational culture in two Awarding Organisations.*

Stuart Shaw (Cambridge Assessment International Examinations): *The CEFR as an assessment tool for learner linguistic and content competence: assisting learners in understanding the language proficiency needed for specific content goals in the CLIL classroom.*

Stuart Shaw (Cambridge Assessment International Education), Victoria Crisp (Research Division) and Sarah Hugues (OCR): *A framework for describing comparability between alternative assessments.*

Tim Oates, co-authored with Philippa Griffiths (Research Division): *The 'grey history' of assessment: understanding the origins of England's new model of assessment of practical work in Science.*

Sylvia Vitello and Carmen Vidal Rodeiro, co-authored with Lucy Chambers (Research Division): *Moderation of non-exam assessments: a novel approach using comparative judgement.*

Irenka Suto convened a symposium on the innovative research area of errors in examination papers: *The rare but persistent problem of errors in examination papers and other assessment instruments*. The following three papers were presented and were followed by active discussions and feedback:

Irenka Suto, co-authored with Jo Ireland (Research Division): *'To err is human' but it's time to go deeper. An analysis of human and system level challenges in the construction of assessment instruments.*

Joanna Williamson and Irenka Suto, co-authored with Jo Ireland and Sylwia Macinska (Research Division): *On the psychology of error: a process analysis method for understanding error detection during the construction of assessment instruments.*

Sylvia Vitello and Nicky Rushton (Research Division): *How and why do errors occur? Insights from people directly involved in assessment instrument construction.*

Migration Research Methods workshop

Jackie Greatorex attended the Migration Research Methods Workshop that took place in Cambridge on the 13th of January. She presented a paper on *Intelligence gathering and networking.*

Association for Science Education

Tim Oates attended the Association for Science Education annual conference which took place at the University of Reading in January 2020, and presented a keynote titled *'Learning everything, learning nothing, or learning something from international comparisons of science curricula'*. There is much contention around the extent to which international comparisons can be used for domestic policy development and improvement of practice. His presentation looked at the principles of robust transnational comparisons, and how 'policy learning' can be a legitimate activity, in contrast to naïve 'policy borrowing'. It has focused particularly on recent progress in the development of Science and Mathematics curricula and the insights which can be gained from well-grounded transnational comparisons.

Data Bytes

Data Bytes is a series of data graphics from Cambridge Assessment's Research Division, designed to bring the latest trends and research in educational assessment to a wide audience. Topics are often chosen to coincide with contemporary news or recent Cambridge Assessment research outputs.

The following Data Byte has been published since *Research Matters*, Issue 28:

- December 2019: Popularity of A level subjects among university students.

Publications

The following reports and articles have been published since *Research Matters*, Issue 28:

Constantinou, F., & Chambers, L. (2020). Non-standard English in UK students' writing over time. *Languages and Education (ahead of print)*. Available online at https://www.tandfonline.com/doi/full/10.1080/09500782.2019.1702996

Darlington, E. (2017, circulated in 2020). *What is a non-specialist teacher?* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at https://www.cambridgeassessment.org.uk/Images/562865-what-is-a-non-specialist-teacher-.pdf

Gill, T. (2019). *Progression from GCSE to A Level, 2017*. Cambridge Assessment Statistics Report. Cambridge, UK: Cambridge Assessment. Available online at https://www.cambridgeassessment.org.uk/Images/560531-progression-from-gcse-to-a-level-2017.pdf

Shaw, S.D., & Crisp, V. (2020). An approach to validation: Developing and applying an approach for the validation of general qualifications. Research Matters: *A Cambridge Assessment publication*, Special Issue 3 (First published 2012). Available online at https://www.cambridgeassessment.org.uk/Images/577704-research-matters-special-issue-3-an-approach-to-validation-republished-with-afterword.pdf

Vidal Rodeiro, C. L. & Stuart Shaw, S. D. (2020). The Cambridge Program in the State of Washington: Students' Characteristics, Courses Taken, and Progression to Postsecondary Education. *College & University. Educating the Modern Higher Education Administration Professional,* 95 (1), Winter 2020, 2–17. Available online at https://www.aacrao.org/research-publications/quarterly-journals/college-university-journal/issue/c-u-vol.-95-issue-1-winter-2020

Sharing our research

We aim to make our research as widely available as possible. Listed below are links to the places where you can find our research online.

- *Journal papers and book chapters:* www.cambridgeassessment.org.uk/our-research/all-published-resources/journal-papers-and-book-chapters/
- *Research Matters* (in full and as PDFs of individual articles): www.cambridgeassessment.org.uk/research-matters
- *Conference papers:* www.cambridgeassessment.org.uk/our-research/all-published-resources/conference-papers/
- *Research Reports:* www.cambridgeassessment.org.uk/our-research/all-published-resources/research-reports/
- *Data Bytes:* www.cambridgeassessment.org.uk/our-research/data-bytes
- *Statistics reports:* https://www.cambridgeassessment.org.uk/our-research/all-published-resources/statistical-reports/
- *Blogs:* www.cambridgeassessment.org.uk/blogs/
- *Insights* (a platform for sharing our views and research on the big education topics that impact assessment around the globe): www.cambridgeassessment.org.uk/insights/
- Our *Youtube* channel: www.youtube.com/user/CambridgeAssessment1 contains Research Bytes (short presentations and commentary based on recent conference presentations), our online live debates #CamEdLive, and Podcasts.

You can also learn more about our recent activities from Facebook, Instagram, Linkedin and Twitter.

Contents / *Issue 29* / Spring 2020

Cambridge Assessment
The Triangle Building
Shaftesbury Road
Cambridge
CB2 8EA
United Kingdom

+44(0)1223 553985
researchprogrammes@cambridgeassessment.org.uk
www.cambridgeassessment.org.uk

7149265903

ISSN: 1755–6031