Contents / Issue 26 / Autumn 2018

2 To "Click" or to "Choose"?: Investigating the language used in on-screen assessment: Rushda Khan and Stuart Shaw

9 Articulation work: How do senior examiners construct feedback to encourage both examiner alignment and examiner development?: Martin Johnson

15 Characteristics, uses and rationales of mark-based and grade-based assessment: Joanna Williamson

22 Is comparative judgement just a quick form of multiple marking?: Tom Benton and Tom Gallacher

28 How have students and schools performed on the Progress 8 performance measure?: Tim Gill

38 Research News: Karen Barden
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:

Credits
Reviewers: Matthew Carroll, Lucy Chambers, Filo Constantinou, Victoria Crisp, and Sylvia Vitello, Research Division, Cambridge Assessment
Editorial and production management: Karen Barden, Research Division, Cambridge Assessment
Additional proofreading: David Beauchamp, Research Division, Cambridge Assessment
Cover image: John Foss Images
Design: George Hammond
Print management: Canon Business Services

All details are correct at the time of going to print in October 2018
Research Matters / 26

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

Observing the ebb and flow of contemporary international discussion about curriculum, it sometimes is hard to see scientific discussion rather than political discourse. We should continue to strive to lay down the principles and practices of the science of measurement, particularly as technologically driven change creeps into day-to-day practices. Globally, accountability systems are becoming more evident, and where exam and test results are available, these become a natural focus of attention.

The UK now has nearly three decades of experience of using examinations outcomes as a principal basis of public accountability, and it is right that we continue to try to bring measurement science to this highly politicised area.

The Progress 8 measure, which is the focus of Tim Gill’s research, is a genuine attempt by the UK Government to introduce both greater validity and sensitivity into target setting and performance measurement. It rightly aims to remedy the serious defects of the simple performance measures introduced in the 1990s, and which persisted despite known severe limitations and highly adverse ‘backwash’ effects. The research findings – that schools with lower attaining intake are less likely to perform well on the Progress 8 measure – has led to calls from various organisations to abandon Progress 8 or indeed, abandon accountability measures based on assessment outcomes. Many of the comments miss a key point in measurement science – that specific measures of curriculum attainment (which also function indirectly as weak measurement of engagement and motivation) cannot measure everything. Not all the factors which give rise to the findings in the research are necessarily outside of schools’ control, or immune from the impact of other aspects of supportive government policy. For example, Daniel Hamlin and Paul E. Peterson’s work in the USA shows that hard-bitten schools in deprived areas are beginning to attract higher performing teachers, by changing their ethos and curriculum principles. We need to keep the science in with the politics.

Tim Oates, CBE Group Director, Assessment Research and Development

Editorial

In the first article of this issue of Research Matters, Rushda Khan and Stuart Shaw give a detailed discussion, illustrated by real examples, of some of the issues that need to be considered when preparing on-screen versions of exam questions. I was particularly interested by their observations about the metaphors of working on-screen (files, folders, notepads) and the implications for instructions to examinees such as “write” or “type”. The second article by Martin Johnson, based on his PhD thesis, uses a detailed analysis of inter-examiner communications to show what a tricky task is faced by Team Leaders when giving feedback that potentially criticises the marking of their team members (who are often experienced teachers and examiners themselves).

In the third article, Joanna Williamson considers the advantages and disadvantages of directly grading assessments (as opposed to marking them and then defining cut-points on the mark scale that identify the different grades). This work is particularly relevant and timely as ‘T Levels’ are introduced in 2020 – decisions about how their various components are to be assessed could have a big impact on whether they achieve their intended purpose.

There has been a lot of recent interest in using comparative judgement (CJ) instead of traditional marking to assess students’ work in certain domains, such as essay writing. Because CJ differs from marking in several important respects, it is difficult to compare them on a like-for-like basis. In the fourth article, Tom Benton and Tom Gallacher argue that much of the apparent advantage of CJ (in terms of its value in predicting scores on other assessments) comes from the simplicity of the marking task combined with using a statistical model to iron out differences in severity among markers. One implication is that traditional marking can be just as effective as CJ if it is combined with simpler mark schemes, multiple marking, and statistical scaling of examiners’ marks.

Finally, Tim Gill uses statistical modelling to explore the relationship between Progress 8 scores and variables at school and student level. He concludes that many relevant variables are beyond the control of schools, and hence that it is questionable whether Progress 8 should be used as a basis for comparing them.

Tom Bramley Director, Research Division

1. See https://www.gov.uk/government/publications/introduction-of-t-levels/introduction-of-t-levels
To “Click” or to “Choose”? Investigating the language used in on-screen assessment

Rushda Khan and Stuart Shaw Cambridge Assessment International Education

Introduction

In this article, we consider the extent to which the language used in on-screen examination questions ought to differ from that of paper-based exam questions. We argue that the assessment language in screen-based questions should be independent of the mode of delivery and should focus on relevant and expected test-taker cognitive processing required by the task, rather than on the format of the response. We contend that medium-independent language improves how well a question will measure the knowledge, understanding and/or skills of interest by allowing test-takers to focus on its content rather than on extraneous, potentially contaminating factors such as technological literacy and mode familiarity. The latter factors may constitute potential sources of construct-irrelevant variance and, therefore, pose a threat to how scores awarded to a performance on a question are both interpreted and used.

‘Translated’ questions

With the “inexorable” advance of technology (Bennett, 2002, p.1) and its inevitable impact on the format, content and direction of educational assessment (McDonald, 2002), there is a growing desire to translate traditional paper-based tests into ones suitable for on-screen assessment. But what do we mean by a ‘translated’ test? Do we mean one that mimics its paper-based original and involves the same wording and task on screen and in as close a format as possible to how it appears on paper?

A translated test should, among other things, attempt to maintain the integrity of the specific features of the task or context deemed most likely to have an impact on test performance when replicated on screen. In addition, it must be ensured that the measurement of the intended constructs (skills, knowledge, and understanding) is not undermined by the presence of unnecessary technological demands (Chalhoub-Deville, 2003). In an age of digital literacy (Spirtes & Bartlett, 2012), it is important that the level of technological familiarity is not integral to the construct(s) of interest (Abedi, 2004, Abedi & Lord, 2001, American Educational Research Association, American Psychological Association; National Council on Measurement in Education; & Joint Committee on Standards for Educational and Psychological Testing [U.S.] 2014, p.67)1. At the same time, however, the integrity of the constructs must not be threatened by the need to remove construct-irrelevant barriers to test performance (Sireci, 2008, p.84). (See Huff & Sireci, 2001; Li, 2006; Russell, Goldberg, & O’Connor, 2003, for an overview of the mounting concerns about the potential threats to the validity of computerised tests.)

The language of instructions in assessments

It has long been accepted that the information provided in the question input (the material contained in a given test question) and in the question instructions (aspects of the task which provide structure and guidance on successful completion) should be presented to the test-taker in an unambiguous manner (Bachman, 1990; Bachman & Palmer, 1996, Carson, 2000; Crisp, Swery, Ahmed & Pollitt, 2008, Shaw & Imam, 2013). One source of test-taker anxiety, according to Madsen (1982), is unclear or ambiguously phrased instructions. Examination questions will necessarily draw upon a number of factors deemed most likely to have an impact on test performance. Such factors can influence the difficulty of the task and how test-takers will perform. Given the requirement to make certain inferences on the basis of test-taker performance, it is crucial that instructions to test-takers are both transparent and accessible. Well-written instructions make it clear to the test-taker exactly what is being asked of them by the test procedure and task, the nature of their expected response, any time constraints and, in some cases, how their response will be scored. It is especially important to provide clear instructions for more complex and/or less familiar tasks (Bachman, 1990, p.124). Bachman and Palmer (1996, p.121) propose three indispensable guidelines for test question instructions. Instructions should be:

1. sufficiently simple for learners to comprehend;
2. short enough so as not to take up too much of the test administration time; and
3. sufficiently detailed for learners to know exactly what is expected of them.

Distinguishing cognitive from technical command words

In considering the language of instructions used in examination questions, it is helpful to use a natural categorisation which is shown in Bloom’s 1956 Taxonomy of Educational Objectives (Bloom, Englehart, Furst, Hill & Krathwohl, 1956). Bloom et al. found it necessary and useful to distinguish between command words which relate to the type of question and those which relate to how the test-taker is expected to organise their response. Thus, in the anatomy of a question, there are two types of command words: those which refer to the cognitive process (e.g., “identify”, “predict”, “explain” and “contrast”) and those which

---

1. Though there are those who contend that computer literacy should be conceptualised as a significant contextual factor interacting with the construct measured in a computer-based language assessment (Jin & Yan, 2017).
The quest for medium independence

By medium independence we mean whether a feature of the assessment can be said to make sense regardless of the medium of delivery. There appears to be a relationship between the category of command word and whether the question is medium-independent. Cognitive command words are always medium independent while technical command words are not necessarily so. By way of illustration, in the example "Explain the difference between a metaphor and a simile", the cognitive process of explanation is the same both on paper and on screen. It is only the

The technological fallacy

Technological fallacy refers to an inherent desire on the part of the test developer to introduce change when responding to the challenges of different delivery formats simply because they are different. The impetus for using medium-dependent language when undertaking direct, word-for-word translations from paper to a digital space appears to
be grounded in two imperatives, namely, the accuracy and clarity manifested in the language of question instructions. Both concepts give rise to at least three scenarios when translating.

**SCENARIO 1:** There is no perceived requirement for change as the accuracy and clarity in language remains the same across both modes.

In the first example question illustrated in Figure 1, there are few perceived challenges in language when translating, as the instruction makes equal sense in both modes. Even though the layout, structure and mode of the response differ, the difference is not deemed to be great enough to modify the instructions. Indeed, the focus is on cognitive processing as the command “Name” does not relate to a particular mode of response.

In the first example question illustrated in Figure 1, there are few perceived challenges in language when translating, as the instruction makes equal sense in both modes. Even though the layout, structure and mode of the response differ, the difference is not deemed to be great enough to modify the instructions. Indeed, the focus is on cognitive processing as the command “Name” does not relate to a particular mode of response.

**Paper version**

13. Name two different types of boat mentioned in the text.

- [ ]

**On-screen version**

18. Name two different types of boat mentioned in the text.

- [ ]

---

**Figure 1: Example question 1**

In this example, a medium-independent approach has been taken. This is appropriate given that the cognitive demand is unchanged and that no mode-specific further instructions are needed.

**SCENARIO 2:** The paper-based language is perceived to lack accuracy and clarity in the new medium but a suggested solution involving medium-dependent language may not be ideal.

In the “flytrap” example shown in Figure 2, there are a number of differences between the paper version and its on-screen counterpart relating to presentation and response format. In the paper version, the test-taker must “circle” the correct answer while on screen they must “click” on the radio button next to the correct answer. It is important to consider whether this difference necessitates a change in question language. The only difference in the content of the question is the technical instruction “Circle the correct answer”, which has been translated to “Select the correct answer”. The reason for the change is principally one of accuracy: it would be incorrect to retain the original technical command because circling is not the expected behaviour of a test-taker answering the question on screen.

This is a scenario where the proposed solution is not necessarily the most elegant as it introduces medium dependency. In this scenario, the ideal solution would be to remove medium dependence across both formats and opt for a medium-independent instruction such as “Choose the correct answer”. Arguably, the style of response is not intrinsic to the answer so it should not matter whether a test-taker circles the response or indicates their choice in another way. However, for the medium-independent instruction to be appropriate for the paper-based test, the layout of the question would need to be such that there is no risk of it being unclear to a marker which response the test-taker intended to indicate (e.g., if a test-taker ticked in between the options in the paper-based version of the flytrap question, it might not be clear which response was intended). In this scenario, the ideal solution would be to remove medium dependency across both modes.

In the next example shown in Figure 3, the difference between formats relates to the following instruction: “Write down the modal method of transport for boys”. This has been changed to “Type in the modal method of transport for boys” in the on-screen format, presumably in order to enhance accuracy across mediums. This scenario constitutes one in which the imperative to improve accuracy does not, however, require a change in language across formats. “Write” is appropriate language for screen as well as paper. The discipline of digital usability would suggest that we should employ ordinary language for digital tasks as much as possible. For this reason, much (though by no means all) of the digital terminology is based on metaphors from ordinary language. We refer to computer-based ‘files’,
'folders', 'notepads' and 'recycling bins' which are literally inaccurate but commonplace terminology in a digital landscape. We also use ordinary language for describing digital actions, such as ‘posting a comment’ and ‘writing an email’. This is an illustration of a type of skeuomorphism: the practice of using real world metaphors to create an affordance that increases understanding and familiarity. In this way, counterintuitively, a reliance on the 'correct' technical language can result in an obstacle to comprehension. "Write" would have been clear on screen too. Alternatively, it might have been ideal to remove the technical command completely and the question could have been rephrased as "What is the modal method of transport for boys?" in both formats. This leads to the second reason for introducing technical instruction: the perceived requirement for greater clarity.

The style of the response, in this next example shown in Figure 4, has changed across the paper and screen formats.

On paper, test-takers are expected to pick a word from the list and copy it out, whereas on screen the words are given in drop-down lists for the test-taker to select from. This removes the need for test-takers to transfer the word, slightly changing the nature of the demands of the task. The on-screen version removes the risk that a word might be incorrectly copied over (though incorrect copying should not affect marks given that incorrect spellings of the correct word would be credited. Also, the words are quite different so it should be clear to markers which option was intended). However, the instruction has not only been modified for accuracy but augmented with the technical instruction “from the drop-down list”. Presumably, the intention here is to introduce clarity to the question by explaining the correct mode of interaction. We may ask whether this really is necessary. It may be sufficient to use medium-independent language such as “Choose the correct materials” in order to maintain the same level of demand.

Like the previous example, the next one shown in Figure 5 shows that the style of response has changed between the two modes, but this time, in a more significant way.

On paper, the test-taker must write down the correct words on the label lines, whereas on screen they must drag and drop words from a list into boxes. Again, this slightly adjusts the demands of the task by not requiring re-writing of words and avoids any risk of spelling errors or poor handwriting affecting marks. There is more chance of weaker writing or spelling affecting marks on the paper version of this task (compared to the question in Figure 4) given that several of the options here are somewhat similar (e.g., stamen, stigma and style).

The notion of using medium-dependent language such as “Drag the correct words” may in this case be more compelling. But even in this case, using such words might be counterproductive, as the instruction “Complete the labelling on the diagram using the words below” may be clear enough in both mediums while also being more succinct.

PropONENTS of the medium-dependent approach may argue that, without showing test-takers the technical mode for answering the question, test-takers may not know what is expected of them. However, this makes an assumption about technical literacy that neglects the notion that the visual cue of a drop-down may be sufficient, and more powerful. A ‘clickable’ blank box with an arrow
that becomes highlighted when hovered over gives a visual instruction in a similar way to lines on a page. We would expect that students encountering lines on a page will be familiar enough with paper to not require an additional instruction of "Write your answer here." If this is the case, it may also be worth considering if a test-taker who is familiar with drop-down boxes from their other digital experiences needs specific instructions.

If basic technical literacy is a requirement of on-screen assessment (as we argue it should be), then it follows that visual cues provide an affordance that is an adequate substitution for a linguistic instruction. It is again part of the technological fallacy that differences need to be consciously accounted for through explicit guidance, rather than acknowledged as differences test-takers can intuitively recognise and account for in their approach to a digital experience.

Any requirement to introduce clarity should be in response to a claim that the test-taker may potentially misunderstand the instruction. The most likely way a test-taker may interpret the last example shown in Figure 5, may be that they might try typing directly into the boxes. However, this would result in technical feedback showing the result of the action (e.g., no text appearing), which should prompt them to try dragging instead. We would hope that this would still be a worst-case scenario as test-takers would be, and ought to be, familiar with the notion of 'draggable' words to drop into spaces. To maximise familiarity, design adjustments should first be made to ensure that the interface follows good digital practice. Indeed in this case the draggable words do not look as draggable as they ought to, and it may be clearer to have a 3D shading effect on the words, for example. If pilot use of the testing system or ongoing feedback from schools suggests that some test-takers struggle even after such improvements to the interface, it may be appropriate to encourage schools to use familiarisation activities prior to the test to avoid such misunderstandings.

In summary, if the test-taker’s expected technical behaviour to produce a response is unclear, it is preferable to improve the test using a technical solution (such as modifying the design), rather than an assessment solution (such as modifying the instruction). As we will go on to discuss, the reason for a lack of clarity in a technical context is not necessarily due to poor assessment language but poor technical affordances.

**Scenario 3:** The paper-based language is accurate but there is a choice of response methods and it needs to be considered whether this gives sufficient clarity.

So far we have looked at examples where there is a specific way in which a test-taker can answer a question – by filling in a text box or choosing a single word from a drop-down list. But there is another unusual type of question where the test-taker may choose the technical steps with which they will produce their response. Consider the scenario shown in Figure 6.
In this example, the on-screen version reproduces the paper-based grid but, as is common with on-screen questions, a palette of tools is provided in order for the test-taker to complete the question. There are a number of possible ways in which the test-taker can draw the reflected image, by:

- drawing the points first, then drawing a line to join the points;
- filling in the cells by clicking on each one;
- using the ‘polygon’ tool to create a shape by clicking on every corner;
- drawing the image freehand; or
- using different combinations of the above.

Any of these approaches may result in a very similar outcome, and it may not even be possible to deduce the tools that the test-taker used based on the completed image. Supplying multiple tools may look peculiar on-screen but, in fact, allowing different approaches arguably mimics a paper approach better than having only one tool: on paper, the test-taker can be as flexible as they wish with the method they use to “draw” their response.

In this case, it would seem particularly undesirable to spell out all the different technical tools in the question itself. A basic attempt may be to append an instruction such as “You may use the line tool, the polygon tool, the polygon tool or the shading tool to answer the question”, but this only announces the suite of tools in the toolbar to the test-taker. More importantly, if the desired approach is to train the test-taker how to use the tools, it would perhaps take up a significant part of the test session to explain each tool in turn, and even then the test-taker would probably like to practise first before committing a response. In general, it is hoped that the tools will be user-friendly and intuitive, and not require much practice.

Thus, introducing multiple potential response techniques makes the argument for including technical commands in the instructions more problematic, as simply signposting them may be unhelpful, and explaining them in detail may be unnecessary and detract from the cognitive demand of the question.

The criterion of technological literacy

As we have argued, attempts can be made to reduce medium-dependent language and it may be helpful to think of a question in each medium in order to achieve this. It also seems that one of the reasons test developers may introduce medium-dependent language is because of an assumption that test-takers need technical guidance in order to understand how to answer digital questions.

Generally speaking, we suggest that test-takers are either ‘baseline technically literate’ or they are not, and the test developer’s approach might affect each category of test-taker differently. Baseline technically literate test-takers are those who are sufficiently technically literate to use technology and are familiar with its conventions. These test-takers are able to navigate to web addresses, recognise and operate scrollbars, open and close windows, and type with confidence. They typically use digital tools on a daily basis for study or leisure purposes. For these test-takers, we may argue that it is unnecessary to provide guidance on how to sit an on-screen test, provided that the quality of the experience is sufficient to follow good digital usability conventions and, therefore, mirror other digital experiences they are already accustomed to. For this reason also, it might be unnecessary to give them instructions on how to use the scrollbar and navigate between questions. If this is acknowledged, then we would argue that these test-takers would also recognise visual cues related to radio buttons, response areas, and drop-down boxes. Any additional explanation of technical facilities should only occur if they are deemed to be more unusual.

For test-takers who are not baseline technically literate, we can assume that they would benefit most from any technical instruction. However, for a test-taker who is not familiar with, or confident to use, computers, it would be unhelpful to introduce specific commands to highlight visual cues on questions. One reason for this is that it may create inconsistency. If, within a question, we need to provide instructions on how to use the computer, then the scope of this must be carefully considered. Otherwise there is no reason why we would only instruct a test-taker to “Use the drop-down list” when we might also need to ask them to “Pick up the mouse”, “Hover over the scrollbar”, “Press the mouse button”, and so on. Additionally, augmenting questions with technical instructions has a direct impact on fairness in a timed exam – time that should be spent on responding to questions is instead spent on learning how to use the technology.

It is likely that there are features of an on-screen test that even baseline literate test-takers would not be accustomed to at all because there is simply not enough of a precedent in their other digital experiences in order to be confident. We have already seen one example of this in Figure 6 where it is unlikely that all students will have encountered the tools they need to use in drawing the response. Another example is if the test-taker needs to write complex mathematical notation using a bespoke toolbar or LaTeX commands. In these cases, an appropriate approach if using on-screen assessments would be to carry out training prior to testing, for example through a familiarisation activity using the tools or copying mathematical notation. This leads to the conclusion that all digital test-takers need to meet the criterion of technical literacy. Those who take an on-screen test should be baseline technically literate through their prior digital experience. We also need to set boundaries on what counts as familiarisation (and therefore sits outside the test) and what is permitted to take up valuable test question landscape. If the criterion for technical literacy is met ‘outside the test’ as we would recommend, then test-takers will be able to use the assessment for its intended purposes.

There is one caveat which we have thus far made passing reference to but which ought to be emphasised. It is that the recommended approach increases the burden on the interface developer to ensure that the correct conventions are used in order to maximise usability. For our argument to hold, a drop-down box should look like a drop-down box, and clicking on a single multiple choice option should show one option clicked rather than two. It also requires the interface to provide adequate technical feedback for a test-taker’s actions: a state change on hover, the highlighting of a word, and displaying the word in situ when it is clicked. If an interaction is not intuitive due to bad design, it is likely that more familiarisation activities or training will be required, even for baseline technically literate test-takers. This would make prior familiarisation time-consuming and frustrating for test-takers if they have to ‘unlearn’ good digital practice.

3. A standard framework for writing mathematical notation on screen.
Quality in a digital landscape has its own guidelines and parameters and a good on-screen assessment should aim to maximise digital usability as well as assessment quality. This can be challenging. Assessment experts are not always digital experts, and vice versa, so there is an inevitable issue in establishing that all the appropriate parameters have been met. We can go some way to address this by acknowledging integrated technical and assessment expertise as essential during the construction of an on-screen test: both types of expertise need to contribute to its quality assurance and sign off.

The criterion for technological literacy is fundamentally related to the notion of expectation. A suitable test-taker for an on-screen test ought to expect a certain mode of response on each question, either through their general digital familiarity (in which case the test interface is responsible for reflecting their other experiences), or through bespoke familiarisation and training (in the case of a new or unusual digital experience which cannot ordinarily be expected). Unlike the test itself, the familiarisation activities are responsible for showing the test-taker how best to use the given technology.

**Conclusion**

In this article we have explored arguments in support of medium independence in assessment language and recommended a number of key approaches:

- When translating from a paper-based assessment to an on-screen assessment there should not be an automatic translation of cognitive command words to technical ones, or an unconsidered appending of cognitive commands with technical commands. Further research is required, however, to verify such assertions.

- If a command in a translated question does not make sense on screen, it is likely that a technical command has been used on paper. The test developer should consider replacing the technical command with a cognitive command in both cases (or at least using a cognitive command on screen).

- Criteria should be set by awarding bodies for the test-taker to be baseline technically literate before sitting an on-screen test.

- An evaluation of each proposed digital assessment needs to be undertaken to assess whether all features of the digital interface ought to give rise to the correct expectation of a baseline technically literate test-taker. If this is not the case, this could be due to the kind of functionality that:
  - is fact common but it has been presented in an unusual style; or
  - is uncommon and is bespoke to a new type of on-screen assessment functionality.

If (a) is the case, efforts should be made to follow best practice of digital convention. If this is not possible, or if the likely scenario is (b), then those specific features should form part of familiarisation or preparatory training activities that sit outside the test session.

- On-screen testing sits between assessment conventions and digital conventions and experts in both areas are needed to ensure high-quality assessment.

By way of response to the question raised in the title of this article, a technical instruction like “Click” is an unnecessary modification if the visual cues and technical literacy of the test-taker meet appropriate digital standards. While it is undeniable that a paper-based assessment is different from an on-screen one in a multitude of ways, it is incorrect to say that it is different in every way and that the language used necessarily needs to be different.

Understanding test-taker expectations and following good practice in technology should allow test-writers to focus on the quality of the assessment, without allowing it to yield to the device upon which it is presented.

**Acknowledgements**

The authors would like to thank Victoria Crisp, Research Division, for her comments and suggestions on the earlier drafts of this article.

**References**


Articulation Work: How do senior examiners construct feedback to encourage both examiner alignment and examiner development?

Martin Johnson Research Division

Introduction

This is a study of the marking feedback given to a group of examiners by their Team Leaders (more senior examiners who oversee and monitor the quality of examiner marking in their team). This feedback has an important quality assurance (QA) function but also has a developmental dimension, allowing less senior examiners to gain insights into the thinking of more senior ones. When looked at from this perspective, marking feedback supports a form of examiner professional learning.

This study set out to look at this area of examiner practice in detail. To do this, I captured and analysed a set of feedback interactions involving 30 examiners across three General Certificate of Education Advanced Level (GCE A Level) subjects. For my analysis, I used a mixture of learning theory and sociological theory to explore how the feedback was being used and how it attained its dual goals of examiner monitoring and examiner development.

UK awarding bodies commonly use specialist marking software to distribute digital copies of students’ examination scripts to examiners for marking. This allows Team Leaders to monitor the marking quality of the examiners under their supervision throughout the marking period. As part of this monitoring activity, Team Leaders are also required to give examiners feedback on their marking. This monitoring and remediation function is an important component of an awarding body's mark scheme application.

As well as having a crucial QA function, previous work has suggested that feedback can also be conceptualised as having an expansive developmental potential for the less senior examiners (Johnson & Black, 2012). Expansiveness is a concept that describes how some contexts help new participants in a professional community to gain access to the important knowledge and values that then allow them to go on to become more independent participants in an activity (Fuller & Unwin, 2003). I argue, in line with Beighton, Poma, and Leonard (2015); Dennen (2004), and some situated learning theorists, that this concept of expansion has important links to learning, since a development in the understanding of professional practice in an area is synonymous with learning to be a professional. This expansiveness includes the type and extent of knowledge transfer, the quality of emotional and practical support for participants, and the appropriate alignment of individual objectives.

Rationale for the study

The acknowledged role that Team Leader feedback has in marking QA processes means that examiner communication is an important area of study. This is particularly the case because of its role in the alignment of Team Leader and examiner thinking which forms the basis of common mark scheme application.

Despite this acknowledged importance, the study of examiner feedback practice is, at present, a relatively under-researched area. This lack of research is the result of a number of specific factors. One factor is that e-feedback practice is still an emerging area of communication, with professional behaviours being inevitably linked to the affordances of the digital marking environments that have recently been adopted across the assessment sector. Another factor links to the challenges of capturing and analysing information that is distributed between individuals across a diverse set of communication channels.

Theory

Learning and communication research suggests a number of potential issues that make the careful study of feedback practice very pertinent.
Interpersonal activity anticipates that examiners develop their understanding of a mark scheme through sharing their (sometimes contrasting) perspectives about how to mark a particular candidate response with their Team Leader. In developing this shared view, examiners draw on shared resources (such as mark scheme documents or candidates’ exam scripts) that invoke concepts that they believe to reside within the cognition of each other.

A consequence of this theory is that where common ground in feedback communication is weak, it is possible that communication will break down, that examiners will fail to establish shared understandings, and that less senior examiners will not become full participants in the professional examiner community.

The second area of relevant theory that I draw on links to the notion of Articulation work (Strauss, 1985). This concept describes how communication helps to coordinate individuals whose work is professionally interconnected. This form of work is carried out by managers to ensure that those around them complete their own tasks, and thus ensure that mutually important strategic goals are attained.

This theory has implications for research methods because it draws attention to the importance of evidencing the minutiae of the professional behaviour that participants carry out, and which often go unnoticed as they are generally taken for granted.

Bringing together the two areas of theory, my study considers the articulation work that Team Leaders carry out through their feedback communication as they build and maintain common ground with the examiners in their team.

**Method**

My study focused on three GCE A Level subjects (Chemistry, Economics and Geography). These subjects were chosen because they included scripts that incorporated subjective items. These items tend to invite performances that require higher order skills, and inevitably involve intricate decision-making on the part of an examiner when applying the mark scheme. Such items are considered to be the most complex item type, and they tend to result in lower levels of examiner agreement
Outcomes

Content and agreement

One of the principal outcomes of my analysis was to be able to identify the types of content that were included in feedback messages. Figure 2 outlines the five types of information that were most commonly found (in descending order of prevalence):

Whilst these findings allow insight into what information contributes to the alignment of examiners’ thinking (e.g., shared information about where and why marking credit is found in a performance, or how examiners are expected to use the specific marking software), this data is only one part of an interesting story.

An important point to consider is that feedback information is generally shared within the broader context of explicitly or implicitly stated disagreement between the participants. Analysis was carried out to measure the relative balance of agreement or disagreement.
Examiner familiarity and experience

My analyses showed that the characteristics of examiner prior experience and familiarity had an impact on the types of feedback given. These outcomes suggest that content and relationship management were entangled in the process of feedback giving.

Using word count as an indication of content, I found that there was significantly more feedback communicated between Team Leaders and new examiners than between Team Leaders and experienced examiners. Messages to new examiners contained on average 116 words, whilst messages to experienced examiners contained on average 75 words.

In addition, and although the groups were highly overlapping, there was significantly more feedback communicated between Team Leaders and unfamiliar examiners than between Team Leaders and with familiar examiners. Unfamiliar examiners were those who had not previously worked with the Team Leader. Messages to unfamiliar examiners contained on average 101 words, whilst messages to familiar examiners contained on average 80 words.

These analyses suggested that Team Leaders were targeting and adapting their communication to the needs of their examiners in different ways, with new and unfamiliar examiners receiving more feedback than other examiners.

Distancing strategies

To make sense of these discrepancies, I analysed any differences in the nature of the information that was being conveyed between these different groups of examiners. The clearest difference was in the way that Team Leaders employed distancing strategies with new and unfamiliar examiners (compared with other examiners).

In my study, distancing strategies refer to the deployment of politeness in discourse. It has already been noted that the presence of negative information in social interaction, such as criticism, disagreement, and interruption, embroils issues of face management. Importantly, it has been observed that politeness can minimise face threat (Goffman, 1967; Morand, 2000), and has been found to be used in professional contexts where bad news needs to be delivered (Sussman & Sproull, 1999).

Theorists also observe that politeness may be of a positive or a negative variety, with each affording the user the opportunity to either increase or reduce the perceived social distance in interaction (Brown & Levinson, 1987). Positive politeness reduces the threat to the recipient’s positive face by accentuating empathy and common ground between the participants, therefore acting as a kind of social accelerator. These positive politeness tactics include admiration (e.g., “I like that way that you approached that problem”) and the use of ‘in-group’ speech forms (e.g., the use of ellipsis and the inclusive pronoun form “we”). On the other hand, negative politeness avoids imposition on the recipient’s negative face (i.e., the desire to act unimpeded) by creating a respectful distance between the participants. Negative politeness tactics act as a form of social brake (Culpeper, 1996) through the judicious use of words to construct messages that include apology, verbal hedging, and honorific term use.

My analyses showed that distancing strategies were used more frequently with new examiners and with unfamiliar examiners. Over the course of their feedback discourse, on average, new and/or unfamiliar examiners’ messages contained around 27–33 distancing strategies, compared with a range of 21–26 distancing strategies for experienced and/or familiar examiners’ messages.

Careful analysis of the feedback discourse showed that distancing strategies were deployed in a number of ways across the feedback corpus. These included the personalisation of messages, the use of apology, and the use of modal forms.

Greetings and closings

Nearly all of the feedback messages included a greeting and a farewell statement. Analyses suggest that the participants used these openings and closings in a purpose driven way so as to achieve particular effects. For example, one Team Leader, (Roy) explained how in general he preferred concise, targeted message writing: “I don’t need to waste [words], the potency of the message goes in the more words you use in my opinion”. At the same time, he rationalised how the effort expended on personalising messages, including elements such as informal greetings and closings (Figure 3), may have a motivating impact on this particular examiner.

When discussing this feedback message, the Team Leader reported: “I have been at the other end of this and you really just want to know where you have gone wrong… I honestly don’t want to dishearten [Eric] To be fair the message was “You are out”. There is a human interaction here. Whereas actually when you are giving just very straightforward feedback “This is right, this is wrong”, you don’t need as much as that, but you need to be a bit softer [here] I guess.”

Apologies

The use of apology is a strategy for manipulating the perceived social distance in interaction. The feedback extract in Figure 4 shows how apology is used by an examiner (Teresa) to preserve the negative face of the Team Leader (Serena) and to reduce intrusion into their professional space.

When discussing this feedback message, the Team Leader reported: “Sorry Serena another question, Q8 p10 - is this enough for L3B2? - thanks Teresa”.

This pattern is also noticeable in the telephone extract in Figure 5, where a Team Leader (Ben) has to inform an examiner (Gerry) that he has to send some standardisation scripts back to the examiner for reconsideration.

This feedback Ben manages engagement through underplaying the seriousness of the disagreement (line 005: “little bit”). The use of apology also reinforces the dispreference related to giving bad news (lines 003, 004, 008, 010: “sorry”, “unfortunately”, “afraid”). Ben then shifts the focus of the conversation (line 013), emphasising (now) before Gerry can dwell on the disagreement. This is a shift towards positive help following the delivery of negative news.

2. All names have been changed to preserve participant anonymity.
A closer look at the use of apology also gives insights into how the participants maintained order and ongoing professional interaction in a context where face threat was present. Log-likelihood ratio analyses that identify keywords (i.e., words that are used significantly more frequently than others in a discourse) showed that the word ‘please’ was used more by Team Leaders than by examiners. In addition, a search of the whole corpus using the search terms [sorry][apol*] located 142 instances of negative politeness forms (e.g., “Sorry this feedback is a bit lengthy”). Most of these apologies (n=125) took used apologetic utterance. Most of these apologies (n=125) took used apologetic utterance. Of these, 111 were located in context (n=111) took used apologetic utterance. Of these, 111 were located in context (n=111) took used apologetic utterance. Of these, 111 were located in context (n=111) took used apologetic utterance. Of these, 111 were located in context (n=111) took used apologetic utterance. Of these, 111 were located in context.

**Modals**

At times Team Leaders were seen to soften the definitiveness of their judgement through the use of modal forms (e.g., could, may, might). Phrases that use these types of words are sometimes called *hedges* and they express tentativeness and avoid strong statements that may be construed as being confrontational (Lakoff, 1973). In the feedback extract in Figure 6, the Team Leader (Ben) responds to a message from an examiner (Tony) by embedding his comments in the original email wording (indicated in red font). Tony has alerted the Team Leader to an apparent mixed message in the mark scheme (lines 005–008), and then asks for clarification on a marking point (lines 012–013). In his response, Ben’s disagreement is weakly stated (lines 005 –008), and then asks for clarification on a marking point. Ben then adds tentativeness and avoids strong statements (e.g., “I am afraid so – possibly a little generous.”). Ben also softens the definite nature of the responses through the use of modals on lines 010, 011, (“may”, “if”), which reduces the implication that the examiner is completely incorrect.

**Discussion**

Through my close analysis of Team Leader and examiner feedback discourse, I have been able to gain insight into the nature of the communication that supports distributed marking processes. Analysis of feedback content shows that examiners are given important information that steers their practice. This communication content helps examiners to refine their interpretations of mark schemes and helps to reduce any marking discrepancies between examiners and more senior examiners/Team Leaders. This content can be interpreted as being a component of an expansive learning environment since it gives new examiners access to the important knowledge and values, that then allows them to go on to become more independent markers.

This analysis also draws attention to the intermental nature of professional development, with examiners developing their understanding of a mark scheme through receiving (sometimes contrasting) perspectives from their Team Leaders on a shared marking performance. Feedback content frequently focused on the location of, and the rationalisations for, marking credit. Drawing on learning communication literature, this content can be interpreted as providing the foundation for the Team Leaders and the examiners to develop a shared view. According to this perspective, productive learning communication relies on the participants developing and maintaining common ground through their discourse.

As well as providing empirical evidence of the transactional content of feedback information, my analyses also give insight into the allied interactional dimension of communication. My analyses show that Team Leaders deploy politeness in their feedback communication in a targeted way: I then explain, using theory, why the common ground that is established through feedback interaction, and the expansiveness that is derived from it, is potentially threatened by the prevalence of negative information (marking disagreement) within the communication. Analysis suggests that the structure of Team Leader feedback communication is influenced by the nature of the information conveyed within the messages. Moreover, interview data showed that this structuring is to some extent conscious and purpose driven on the part of the Team Leaders. Feedback information that conveys disagreement is a negative basis for establishing productive, ongoing relations. Team Leaders appear to structure negative feedback messages in ways that attempted to maintain productive engagement through reinforcing an examiner’s sense of professionalism. This is most clearly demonstrated in the prevailing use of negative politeness strategies in such messages. This is particularly the case with new and/or unfamiliar examiners, with whom Team Leaders would be expected to have the weakest common ground.

Drawing on sociological theory, this form of relationship management through feedback can be interpreted as a form of *Articulation work*. This is a form of ‘taken for granted’ coordination work that ensures that mutually important strategic goals are attained. Team Leaders use feedback to communicate important content to examiners whilst also mitigating the threats to common ground building that pertains to the negative information that the messages sometimes need to convey. The use of negative politeness helps the participants to maintain a respectful professional distance, and a corollary of this is that marking work is maintained (and not curtailed prematurely due to a lack of *examiner will rather than examiner skill*). Having an ongoing feedback interaction over time allows a virtuous cycle of examiner development to be constructed. Ongoing marking experience leads to attendant feedback, a process of...
examiner reflection, and the consolidation of examiner thinking that is reinforced by a Team Leader’s perspective.

The insights from this study set out the complexity of the feedback-giving task, and how it interacts with the nature of professional examiner development. It also gives insights into the nature of the relationships that foster professional development, and the importance of the forms of communication that lay the foundations for both examiner learning and the completion of marking tasks to a high standard. My analyses illuminate the way that Team Leaders manipulate the perceived social distance within their remote feedback communication so as to attend to the dual functions of (a) monitoring the standard of examiner marking, and (b) giving examiners information that supports their ongoing development. These insights could be used to inform any future training that is given to Team Leaders in preparation for feedback-giving practices.

References


Characteristics, uses and rationales of mark-based and grade-based assessment

Joanna Williamson  Research Division

Introduction

Qualifications currently available in England include examples of both mark-based and grade-based approaches to assessment. Mark-based approaches require assessors to assign numerical marks to candidates’ work, assisted by a mark scheme. Grade-based approaches avoid marks altogether, and require assessors to assign grades by evaluating candidates’ work against grading criteria. These are statements that set out the knowledge or skills that must be demonstrated to gain a particular grade (e.g., “Delivers sports/activity sessions using effective communication” in a Sports qualification).

The choice between a mark-based or grade-based approach has important implications for a given qualification’s overall assessment model. The choice determines what kind of judgement assessors are required to make, and the standards they must understand and apply. Mark-based and grade-based approaches also facilitate different ways of combining or aggregating judgements. For example, in a mark-based approach, the marks given for individual tasks can be added together. Despite these significant consequences, little has been written about the rationale for deciding upon a mark-based or grade-based approach, or the consequences of the decision.

This article outlines the characteristics, uses and rationales of mark-based and grade-based assessment approaches. It focuses on the potential strengths and weaknesses of mark-based and grade-based approaches for internal assessment 1 in vocationally-related qualifications (VRQs). Qualifications of this type that are available in England, such as Applied Generals and Technical Awards, include examples of both mark-based and grade-based approaches, and, for this reason, are a particularly interesting context to consider.

Where are mark-based and grade-based approaches used?

Before reviewing the characteristics of mark-based and grade-based assessment approaches in depth, it is worth noting the contexts in which they are used. Table 1 shows where mark-based and grade-based approaches are used within common qualifications taken in England. General qualifications (Grades) in England such as the General Certificate of Education Advanced Level (GCSE A Level) are overwhelmingly assessed using mark-based approaches. Applied and technical qualifications, by contrast, include examples of both mark-based and grade-based assessment.

### Table 1: Common qualification types in England and their assessment approaches

<table>
<thead>
<tr>
<th>Qualification type</th>
<th>Qualification Unit/component</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applied/Technical</td>
<td>BTEC Nationals Level 3</td>
<td>Grade-based</td>
</tr>
<tr>
<td></td>
<td>Internally assessed units</td>
<td>Grade-based</td>
</tr>
<tr>
<td></td>
<td>Externally assessed units</td>
<td>Mark-based</td>
</tr>
<tr>
<td>Applied/Technical</td>
<td>Cambridge Nationals</td>
<td>Mark-based</td>
</tr>
<tr>
<td></td>
<td>Internally assessed units</td>
<td>Mark-based</td>
</tr>
<tr>
<td></td>
<td>Externally assessed units</td>
<td>Mark-based</td>
</tr>
<tr>
<td>Applied/Technical</td>
<td>Cambridge Technicals</td>
<td>Grade-based</td>
</tr>
<tr>
<td></td>
<td>Internally assessed units</td>
<td>Grade-based</td>
</tr>
<tr>
<td></td>
<td>Externally assessed units</td>
<td>Mark-based</td>
</tr>
<tr>
<td>Applied/Technical</td>
<td>General National Vocational Qualification (GNVQ)</td>
<td>Grade-based</td>
</tr>
<tr>
<td>Applied/Technical</td>
<td>NCFE V Certs Level 2</td>
<td>Grade-based</td>
</tr>
<tr>
<td></td>
<td>Internally assessed units</td>
<td>Grade-based</td>
</tr>
<tr>
<td></td>
<td>Externally assessed units</td>
<td>Grade-based</td>
</tr>
<tr>
<td>General</td>
<td>GCE A Level</td>
<td>All</td>
</tr>
<tr>
<td>General</td>
<td>GCSE</td>
<td>All</td>
</tr>
<tr>
<td>General</td>
<td>Cambridge Pre-U</td>
<td>All</td>
</tr>
</tbody>
</table>

Framework for considering strengths and weaknesses of mark-based and grade-based approaches

Neither mark-based nor grade-based assessment approaches are inherently ‘better’. To judge their strengths and weaknesses, they need to be evaluated in terms of how well they support a particular assessment purpose. Broadly speaking, the assessment purpose of most qualifications is reliable and valid assessment of specified knowledge and skills, that is acceptable to those involved: teachers, students, regulators, awarding bodies and employers. More specific assessment purposes, however, depend on the particular aims and context of the qualification (e.g., to certify competence in a particular occupation).

Empirical evidence on the reliability of internal assessment, particularly in VRQs, is rare (Johnson, 2013). However, a broader body of research offers evidence about factors that affect reliability more

---

1. In external assessment, an awarding body sets the assessment task(s), specifies when and how candidates take the assessment, and carries out the marking/grading. In internal assessment, one or more of these activities is carried out by the candidate’s school/college or workplace instead of an awarding body.

2. Many qualifications, including those listed in Table 1, are made up of multiple units or components. Candidates are assessed in these units/components separately, and the overall qualification result is then determined by combining the unit-level results. Different units/components may assess different areas of knowledge and skill. For example, a Sports qualification might contain three units/components assessing Sport Performance, Physiology and Coaching.
generally. Marker agreement can be increased either by improving the expertise of assessors, or by making the marking/grading task less demanding (Suto & Nádas, 2008). Since neither mark-based nor grade-based approaches can directly improve assessor expertise, the way in which they can best support reliable internal assessment is by reducing the demand of the marking/grading task.

To understand the marking/grading task demand in internal VRQ assessment, it is important to acknowledge two important aspects of the marking/grading context. The first is that internal VRQ assessments frequently assess portfolios of evidence from tasks such as projects and practical activities, characterised by relatively low levels of task constraint (Ahmed & Pollitt, 2011). Assessing portfolios is arguably fairly demanding, due to “the complex nature of the assessment task (multiple, often non-standard and probably complex assignments)” (Johnston, 2004, p.395) and the fact that assessors must deal with a large amount of information (Johnson, 2008a, p.28), which may not necessarily be well-organised. The second aspect to acknowledge is that, internal assessors typically form a much larger group than examiners, but may only have experience of assessing their own students, and may have varying levels of skill.

Characteristics of mark-based and grade-based approaches

Considered in abstract, the only inherent difference between mark-based and grade-based approaches is the difference outlined at the start of this article. In a mark-based approach, assessors assign numerical marks to candidate work, assisted by a mark scheme, whereas in a grade-based approach, assessors evaluate candidates’ work against grading criteria to decide upon a grade, avoiding marks altogether. Mark-based and grade-based assessments found in practice, however, differ across a range of characteristics. Some differences result directly from the choice of a mark-based or grade-based approach, whilst others instead reflect the rationales and characteristics of assessment traditions that are strongly associated with mark-based or grade-based approaches. The differences can be classified into those to do with making judgements, and those to do with aggregating judgements. The following sections explore these differences, and the evidence for how they may form advantages or disadvantages.

Making judgements

Rating scale

A mark scheme designed for tasks with low levels of constraint, such as those found in internal VRQ assessments, typically assigns marks by describing different levels of candidate performance. Figure 1 shows an extract from such a levels-based mark scheme. Each row describes a different level of performance, and the mark bands show what ranges of marks should be awarded to work at these levels. Using this mark scheme, assessors compare candidates’ work to ‘the standard’ by deciding which level best describes the candidate’s work, and to what extent. Levels of candidate performance are discriminated through comparison to marking criteria that describe the quality of candidate work expected at different points on the mark scale.

<table>
<thead>
<tr>
<th>Level</th>
<th>Mark range</th>
<th>Marking criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band 1</td>
<td>1-4 marks</td>
<td>Selects formulae and functions to produce a solution which has limited capacity to meet user requirements.</td>
</tr>
<tr>
<td>Band 2</td>
<td>5-7 marks</td>
<td>Selects formulae and functions to produce a solution that includes elements of efficiency and satisfies some of the user requirements.</td>
</tr>
<tr>
<td>Band 3</td>
<td>8-10 marks</td>
<td>Selects formulae and functions to produce a solution that is effective and efficient and in the main accurately meets user requirements.</td>
</tr>
</tbody>
</table>

Figure 1: Extract from a levels-based Cambridge Nationals mark scheme (OCR, 2017)

In grade-based approaches to internal VRQ assessment, candidate work is typically compared to the standard through the assessor deciding whether individual grading criteria have been met or not met. Individual grading criteria specify separately identifiable skills or competencies, and candidates are assessed against each. This is not a necessary feature of grade-based approaches, but is the dominant practice found in grade-based VRQ assessments, including BTECs and Cambridge Technicals.

The key advantage of a mark-based approach is that it can capture more information about candidate performance, recording the extent to which a candidate achieved something, not just that it was (or was not) achieved. A commonly found disadvantage of mark-based approaches is that it can be difficult to differentiate between multiple points on a rating scale, even for experienced assessors (Johnson, 2011). This increases task demand, and may also lead to underuse of certain marks, with a negative effect on reliability.

Wolf (1993) discusses rating scale choice in the specific context of criterion-referenced assessment, emphasising that the choice depends on “the inherent logic of the subject (whether there are key, recognisable, thresholds) and also on the context (the use to which it is being put)” (p.23). In particular, Wolf stresses:

> There is nothing about criterion-referenced testing which ties it to a pass-fail, on-off approach. Criterion-referenced assessments produce a ‘distribution’ of performance (or, in a formal test, marks) in exactly the same way as any other assessment does. A single pass-fail is ONE way to partition that distribution but only one. (p.23)

Wolf argues that a binary (met/not met) judgement may not be the most reliable way to assess performance against a standard, on the basis that qualifications assess human behaviour, and “...human behaviour does not usually fall into ‘on/off’ categories but instead along a continuum” (Wolf, 1993, p.25). For this reason, “The decision about whether someone should be placed on one side or the other of a particular criterion line can consequently be difficult to make – and the assessor’s judgement fallible” (p.25). The consequences for reliability are that aggregation of binary decisions may lead to a result that does not reflect candidate achievement: the result “may be quite untrue to the underlying performance of the candidate”, whereas the aggregation of more finely-graded judgements can be fairer because of the additional information that is captured at the first assessment stage (p.25).

Mark scheme structure

Levels-based mark schemes vary in the extent to which they are holistic or analytic in structure. In holistic levels-based mark schemes (see, for
example, Figure 1 in the Appendix], each level of response described may encompass multiple component skills or features, without guidance on how to weight their relative importance. Assessors are required to decide which band provides the best fit to the candidate’s response, although a candidate response may exhibit varying levels of performance against the component skills described. The disadvantage of such ‘best fit’ levels-based mark schemes is that marker agreement may be undermined by leaving aspects of the assessment judgement without explicit guidance. Centres and assessors may vary in how they weigh component skills, which introduces a threat to validity.

An analytic levels-based mark scheme (see, for example, Figure A2 in the Appendix), describes levels of candidate performance separately for multiple strands or component skills. Ahmed and Pollitt (2011) argue that explicitly analytic mark schemes are the most reliable way to mark unconstrained tasks, because they minimise the chance that an assessor has to make an assessment judgement without guidance from the mark scheme. Explicitly analytic mark schemes give “more help … by making it clear what distinguishes better from poorer responses” (Tisi, Whitehouse, Maughan, & Burdett, 2013, p.24). This is particularly valuable for tasks that are unconstrained, with hence less predictable candidate responses. A disadvantage of analytic mark schemes, however, is that, the more closely component skills are analytically decomposed and specified, the more information assessors are required to assimilate, and the more the mark scheme resembles a lengthy points-based mark scheme instead of a levels-based mark scheme (Pinot de Moira, 2013).

Points-based mark schemes allocate marks to objectively identifiable words or ideas that are listed in the mark scheme, and are considered inappropriate for tasks with low levels of constraint (such as those found in internal VRQ assessments), not least because they require the pre-specification of all possible credit-worthy points.

Assessment criteria

The clarity of assessment criteria is a direct way for a marking or grading approach to affect the demand of a marking/grading task and thereby marker agreement. Ahmed and Pollitt argue that assessors are most effectively guided when a mark scheme “interprets the important aspects of the trait in the specific ways in which they should appear in responses” (Ahmed & Pollitt, 2011, p.275). Even so, the interpretation of descriptive assessment criteria can present difficulties. These arise from the need to interpret the vocabulary used, which will be influenced by subject-specific construct perceptions, assessors’ professional experiences, and also their personal experiences and values (Johnson, 2013; Nádas, Suto, & Grayson, 2012). Assessors may encounter difficulties in separating quality from quantity when making judgements, and may perceive some criteria to overlap. The findings of Johnson (2008a, 2008b) illustrated these difficulties in the context of mark-based VRQs, and showed that marker agreement was negatively affected where these difficulties occurred.

Grade-based approaches found in internal VRQ assessments, such as those of Cambridge Technicals and BTec qualifications, are underpinned by the logic of an ‘objective’ judgement of performance against individual grading criteria. Such grade-based approaches typically feature grading criteria that are far more concise than descriptive marking criteria. Taken in isolation, this seems to be an advantage since it imposes lower cognitive demands than multiple pages of high-density text (Pinot de Moira, 2013). On the other hand, the much shorter criteria necessarily provide assessors with less information. Furthermore, whilst descriptive criteria in a mark-based approach are notoriously vulnerable to variable interpretation, the concise criteria in grade-based approaches may still suffer from this problem.

Grade-based assessment has in the past been assumed to be more reliable than other forms, due to the ‘transparency’ of criteria based on competences. In practice, research and experience have conclusively demonstrated that “even tightly written specifications of criteria are capable of multiple interpretations” (Baird, Beguin, Black, Pollitt, & Stanley, 2012, p.55; Eraut, Steadman, Trill, & Parkes, 1996; Greatorex & Shannon, 2003; Wyatt-Smith & Klenowski, 2013). The extremely detailed criteria written for GNVQ assessment, for example, could not prevent “very low agreement on key indicators” and a “‘lack of consensus’ relating to grading standards” (Wolf, 1998, p.438). Carter and Bathmaker (2017), more recently, illustrated how supposedly unambiguous assessment standards in grade-based VRQ assessment can be interpreted by assessors as uncomfortably flexible.

Wolf (1993) argues that it is misleading to conceive of competency and criterion-based assessment approaches as qualitatively different from other assessment approaches, since no criterion can ever be entirely transparent, and human performances are intrinsically variable. It follows that candidate performance “cannot be fitted mechanistically to either a written list of criteria or an exemplar” and will require at least some assessor judgement (Wolf, 1993, pp.16–17). Wolf emphasises that “while assessment systems may vary in the degree to which these complex judgements come into play, such judgements are universal to all assessments” (Wolf, 1993, p.17).

In both mark-based and grade-based approaches, efforts can be made to increase the clarity of assessment criteria. However, there are risks attached to doing so. One is that criteria containing more clarifying information for assessors (with the aim of lowering marking task demand) almost invariably add to the total amount of information that must be read and understood. Pinot de Moira (2013) concluded that criteria giving assessors less information could result in more reliable marking due to the lower cognitive demand when mark schemes were uncluttered and required less reading. A second risk is that by pursuing clarity of criteria in order to increase the reliability of assessment, criteria can be altered in a way that threatens validity. Eraut et al. (1996, p.5) summarised the tensions between clarity of standards and validity in vocational assessment as follows: “The search for perfect reliability leads toward tests whose completion and marking allow for no possible margin of error and end up measuring nothing worth knowing. Pursuing perfect reliability leads to meaningless assessment.”

Aggregating judgements

Mark-based and grade-based approaches may differ in both the aggregation of assessment judgements within units/components, and in the aggregation of unit-level results to form the overall qualification result. Since all aggregation involves loss of information, it can be argued that the optimal aggregation of assessment judgements would be no aggregation at all. Reporting a lengthy profile of marks or grades, however, tends not to be acceptable to users of results (such as employers and universities).

Within mark-based assessment, two main aggregation methods are possible. Either marks can be added together, or marks can be mapped onto a different numerical scale and then added. An example of the
latter approach is the Uniform Mark Scale (UMS) used until recently in modular CCE A Levels (Gray & Shaw, 2009). Within grade-based assessment, many methods of aggregation are possible. One approach is to convert grades into points, and then calculate a total score or average, for example, Grade Point Average (GPA). Another approach is to apply an algorithm or rule based on the profile of unit-level grades. For example, the rule could specify that the qualification grade is determined by the lowest grade achieved on any single unit.

The methods used to aggregate judgements can lead to very different outcomes from the same initial set of results. Aggregation affects assessment reliability, and also affects the validity and acceptability of a qualification result, due to decisions about what the aggregation preserves and what is lost. Thomson (1992), discussing possible aggregation methods for modular qualifications like the General Certificate of Education (GCSE), emphasises that different aggregation methods must be evaluated in light of the purpose of the qualification: “...the philosophy of the awarders is an essential element in trying to decide which system is best, or most appropriate. Whether the system is quantitative or qualitative, the decision to choose between them will always be qualitative” (Thomson, 1992, p.7).

The characteristics of aggregation methods that differ include the degree of compensation allowed between units, the reliability of overall results, the likelihood of anomalous results, and the loss of information that occurs between assessment and overall grades. The following sections compare mark-based and grade-based aggregation methods in these terms, and consider their advantages and disadvantages for supporting common assessment purposes.

Compensation
Mark-based approaches to assessment almost invariably involve the addition of marks achieved for different questions or tasks, and hence an element of compensation. If a unit-level result is obtained by adding the marks achieved for each learning outcome, for example, low performance against one learning outcome can be offset by high performance against another. By contrast, many grade-based approaches to assessment, including those used in Cambridge Technicals and BTECs, use conjunctive or hurdle rules to aggregate within units, whereby a candidate must achieve a given grade across all assessment criteria before they are awarded that grade for the unit as a whole.

Compensation is “a guiding principle of most examinations” (Cresswell, 1988, p.370) and in the context of CQs is widely agreed to be fairer and more reliable than non-compensatory approaches (Cadwallader, 2014, p.14; Cresswell, 1987). Arguments for fairness include offering students flexibility, and avoiding results being determined by “the simplest task in which you fail” (Cresswell, 1987, p.251, citing Forrest & Shoesmith, 1985). The latter is considered not just unfair in a general sense, but specifically at odds with the goal of rewarding achievement. Where assessment criteria are differentiated into levels of performance and aggregated using a conjunctive or hurdle approach, there is “failure to reward appropriately the achievement of candidates whose demonstrated attainments do not fit the hierarchical pattern anticipated” (Cresswell, 1994, p.50).

Compensatory aggregation may also be considered an advantage due to matching the practice of CQs. Where students (and their parents and teachers) are accustomed to compensatory aggregation, and qualification results are likely to be compared with results from assessments with compensatory aggregation, assessment hurdles may be considered particularly unfair. In 2018, the Office of Qualifications and Examinations Regulation (Ofqual) advised awarding bodies that the ‘must pass’ examination hurdles present in Applied General and Tech Level qualifications should be reconsidered, since A Level qualifications did not include similar hurdles, and Applied General and Tech Level candidates were therefore being unfairly disadvantaged (Beach, 2018, March 15).

It can be argued that it is problematic to evaluate VRQ assessment practice in terms of GQ practice, since VRQs have different aims and purposes. However, the differences can be overstated. Ofqual’s argument recognises that there exists substantial overlap in the uses of VRQ and GQ results. Not least, many Applied General and Tech Level candidates will use their results to compete with GQ candidates for entrance to Higher Education and employment.

The logic of hurdle-based aggregation reflects the logic of criterion-referencing and competence-based assessment”. Compensation between skills is intentionally avoided, so that a given grade ‘guarantees’ that a particular set of criteria has been met (Creatorex, 2001, p.7), a position that depends on assuming that individual criteria are assessed without error. The guarantee is the principal advantage of hurdle-based aggregation. It becomes “possible to ensure that all candidates have successfully completed tasks deemed appropriate to the award of their grades” (Good & Cresswell, 1988, p.49). The lack of this guarantee is, correspondingly, the chief disadvantage of compensatory approaches. The compensation “obscures the meaning of a grade” (Cadwallader, 2014, p.4), since a student with a given grade may have performed uniformly well across assessment criteria, but may also have excelled in some (unknown) areas whilst having weaknesses in other (unknown) areas.

The disadvantages of hurdle-based aggregation correspond, as expected, to the advantages of compensatory approaches. When used within units, hurdle-based aggregation means that unit-level results are determined by the lowest-graded criterion that a candidate has failed to achieve. This means that students’ grades “no longer reflect their average achievement” (Good & Cresswell, 1988, p.49), and is arguably incongruent with the aim of rewarding achievement. The introduction of hurdles can also lower achievement at qualification level, sometimes dramatically (Baumre & Yoro, 2002, Taylor, Pritchard, & Gray, 2006). Acquah and Malpass (2015), discussing assessment design decisions for a Technical Baccalaureate, suggest two further disadvantages of hurdle-based aggregation. Firstly, hurdles may demotivate learners who feel they are unlikely to pass one element. Secondly, a washback effect could direct centres and students to focus disproportionately on elements that may not be passed, neglecting other areas of learning.

Information loss
Mark-based aggregation methods preserve a high level of information about students’ performances at unit level. Retaining the fine-grained information about unit-level performance “until the last moment” means that overall grades “can be based on maximum evidence” (Thomson, 1992, p.42).

3. Wolf (1995) discusses how competence-based assessment can be seen as a specialised, vocationally-focused form of criterion-referenced assessment, and offers the following definition: Competence-based assessment is a form of assessment that is derived from the specification of a set of outcomes, that so clearly states both the outcomes – general and specific – that assessors, students and interested third parties can all make reasonably objective judgements with respect to student achievement or non-achievement of these outcomes, and that certifies student progress on the basis of demonstrated achievement of these outcomes. Assessments are not tied to time served in formal educational settings (Wolf, 1995, p.1)
Grade-based assessment collects less fine-grained information about candidate performance than mark-based assessment at the point of initial assessment judgements. Hurdle-based aggregation methods within units then preserve only the grade achieved across all learning outcomes. It can be seen as a disadvantage of such aggregation methods, both for perceived fairness and for the usefulness of qualification grades, that a candidate who has performed minimally at a particular grade (i.e., is judged to have ‘just’ met the necessary criteria and no more) is awarded the same grade as a candidate who has met all necessary criteria easily, and has perhaps met a high proportion of higher grade criteria in addition. On the other hand, it could be argued that the level of distinction between candidates is sufficient for the purposes of the qualification. The relevance of grading (at all) in VQs and VRQs can be challenged (e.g., Johnson, 2008c), and although regulated qualifications including Applied Generals and Tech Levels are required to grade and reward achievement (DfE, 2016), this does not necessarily require the fine-grained distinctions enabled by a mark-based approach. It could alternatively be argued that the information loss is significant, but acceptable in light of the overall weight of advantages conferred by a grade-based approach.

As in within-unit aggregation, different methods of aggregating unit-level results preserve different levels of information. A conjunctive approach in which the qualification grade must be met or exceeded in every unit (meaning that the qualification grade is equal to the candidate’s lowest unit-level grade) will discard any information about higher achievements by the candidate in selected units. A points-based system in which unit-level grades are converted to points and then added, on the other hand, will capture these achievements, as well as introducing compensation between units.

Reliability

The compensation in mark-based aggregation confers an advantage in terms of reliability. Adding multiple marks allows compensation between positive and negative errors of measurement (not just between student strengths and weaknesses), so that “aggregate marks can reasonably be taken to be a reliable measure of candidates’ general performances” (Cresswell, 1988, p.363). The absence of this ‘cancelling out’ of measurement error is, correspondingly, a disadvantage of aggregation approaches that do not permit compensation. Chester (2003) noted that where results or measures are combined using a conjunctive aggregation rule, “the reliability of the decision (in the example, whether a student should receive a diploma) is that of the least reliable measure” (p.33).

Douglas and Mislevy (2010) used a simulation method to estimate the classification accuracy of component-based assessments or measures formed by different aggregation rules. Considering multiple measures of agreement, Douglas and Mislevy (2010) found that “adding up scores provides the most consistently accurate decision for all students” whilst “the conjunctive rule [hurdle aggregation] does a better job of correctly identifying students who have not acquired the necessary skills at the expense of misclassifying students who in fact have acquired such mastery” (p.302).

Cresswell (1988) suggested that non-compensatory aggregation is appropriate only when knowledge about performance in individual components is the priority. In such cases, the overall grade is required to carry a different meaning to the meaning it holds in GQs. The argument for the greater reliability of mark-based aggregation no longer applies, since component-level results “do not purport to be and should not be judged as measures of general performance” (p.364), and their aggregated marks do not capture what is supposed to be assessed. In the case of a VQ used to certify competence, the disadvantages introduced by hurdle-based aggregation are accepted, since the risk of a ‘false positive’ (incorrectly certifying competence) is a more serious problem than failing to reward instances of high-achievement. However, VRQs such as Applied Generals, Tech Levels and Technical Awards do not certify competence in an occupation. In fact, it is required that they are not occupation specific and that they reward student achievement (DfE, 2016). A high level of compensation (which risks awarding too generously to students with gaps in their understanding, and obscures the meaning of the grade awarded) therefore seems to represent a lower threat to the assessment purpose than aggregation with little or no compensation (which risks failing to reward achievement in the areas where students have excelled).

Anomalies

The aggregation of unit-level results can result in several kinds of anomaly. Firstly, the aggregation method can cause one given set of unit grades to result in more than one overall grade – as would happen when, for example, units are graded, but unit-level marks are aggregated and final grades are based on the overall aggregated mark scale. Secondly, an aggregation method can result in a ‘mismatch’ between unit grades and overall grade, for example where a profile of unit-level Merit grades results in an overall Distinction grade. Thirdly, an aggregation method can result in multiple sets of unit grades resulting in the same overall grade. Of these anomalies, the third type is not considered problematic (unless prevalent to a very high degree, for example, almost all unit grade profiles resulting in an overall Merit grade). The first two types, on the other hand, are considered undesirable (Cresswell, 1988; Thomson, 1992).

Mark-based aggregation via UMS avoids the second type of anomaly, since the qualification-level grade boundaries are obtained by adding the UMS grade boundaries at unit level. The first type of anomaly does occur, so that “the same grade profile does not lead to a well-defined overall grade” (Taylor et al., 2006, p.20). However, the variation among overall grades is caused by retaining information about the quality of performances at unit level, which, in Thomson’s view, makes such anomalies “justifiable” (Thomson, 1992, p.137), particularly when transparency is ensured by reporting unit-level UMS results.

A chief advantage of grade-based methods is that anomalous qualification-level grades do not occur. Since unit-level grades are themselves the information used to calculate the overall grade, the same set of unit-level grades always results in the same qualification-level grade. A secondary advantage of this is that students are very easily able to assess their progress towards an overall qualification result.

Conclusions

Table 2 summarises the potential advantages and disadvantages of mark-based and grade-based assessment approaches, as used in VRQs. The overall implications of choosing between mark-based and grade-approaches are not straightforward. Although research has identified numerous assessment features which support marker agreement, these
can be in tension with each other (e.g., providing clarification versus conciseness), and it is not clear what the net effects of particular marking and grading approaches are. In the context of VRQs, there are well-documented challenges that apply to both mark-based and grade-based approaches. Studies in VRQ assessment have not explicitly compared the extent to which mark-based and grade-based approaches succeed in addressing these challenges.

Within VRQs, both mark-based and grade-based approaches assess candidates’ performances against assessment criteria. However, there are differences in how they do this. This article has highlighted ways in which these differences reflect different underlying assessment rationales. The features of mark-based approaches to VRQ assessment typically reflect dominant practices from GQs, and an underlying concern with recognising the extent of candidate achievement. By contrast, grade-based assessment in VRQs tends to reflect the strong association of grade-based assessment with criterion-referenced and competency-based assessment.

It is important to reiterate that the advantages and disadvantages of each approach can only be weighed in light of assessment purpose. In the case of VRQs, this may not always be transparent to students and other stakeholders, especially in comparison with better known examples such as GQs and wholly VQs.

### References


### Table 2: Summary of marking/grading approach advantages and disadvantages

<table>
<thead>
<tr>
<th>Mark-based approaches</th>
<th>Grade-based approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ Detailed criteria in level-based grid provide guidance on distinguishing better responses from poorer responses, supporting validity and marker agreement</td>
<td>▼ Grading criteria provide assessors with less information on how to distinguish between responses</td>
</tr>
<tr>
<td>▶ Assessors able to recognise different levels of achievement at the same criteria</td>
<td>▼ Forces assessors to make binary decisions about performance for each criterion</td>
</tr>
<tr>
<td>▶ Detailed descriptive criteria vulnerable to variable interpretation</td>
<td>▼ Concisely written criteria still open to variable interpretation</td>
</tr>
<tr>
<td>▶ High-level marking task demand due to amount of information given to assessors, likely to weaken marker agreement and also acceptability</td>
<td>▶ Brevity of assessment criteria is likely to result in lower marking task demand, which may result in higher marker agreement and help acceptability</td>
</tr>
<tr>
<td>▶ Ambiguity introduced if a levels-based mark scheme describes multiple skills/criterion is unambiguous</td>
<td>▶ Separate (binary) decision for each criterion is unambiguous</td>
</tr>
<tr>
<td>▶ Compensation within units supports reliability of unit-level results</td>
<td>▼ Hurdle aggregation within units is a risk to the reliability of unit-level grades</td>
</tr>
<tr>
<td>▶ Aggregation via addition of marks supports reliability of overall results</td>
<td>▼ Points-based aggregation of units uses all information captured at unit level</td>
</tr>
<tr>
<td>▶ High levels of compensation may be perceived as more fair to students</td>
<td>▼ Hurdle aggregation within units is hard to justify if certification of competence is not the aim</td>
</tr>
<tr>
<td>▶ High levels of compensation perhaps better aligned with aim to recognise student achievement</td>
<td>▼ Hurdle aggregation within units can be seen as against aim of rewarding success</td>
</tr>
<tr>
<td>▶ Performance against criteria cannot be deduced from unit or overall grade</td>
<td>▼ Unit grade is able to ‘guarantee’ that candidate has met minimum level of performance against the criteria for that unit</td>
</tr>
</tbody>
</table>

Legend: ▶ = Advantage ▼ = Disadvantage


---

**Appendix**

<table>
<thead>
<tr>
<th>Level</th>
<th>Marks</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>15–20</td>
<td>Accurate and detailed knowledge of … is demonstrated. The response is highly focused on the question. Evaluation/discussion is effective and logically developed.</td>
</tr>
<tr>
<td>2</td>
<td>8–14</td>
<td>Generally accurate knowledge of … is demonstrated, with some omissions. The response is generally focused on the question. Some effective evaluation/discussion is present.</td>
</tr>
<tr>
<td>1</td>
<td>1–7</td>
<td>Limited knowledge of … is demonstrated. The response relates to the topic but addresses the question only to a limited extent. Response is mainly descriptive. Evaluation/discussion is limited or absent.</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>No relevant content.</td>
</tr>
</tbody>
</table>

**Figure A1**: Holistic levels-based mark scheme – hypothetical example

<table>
<thead>
<tr>
<th>Level</th>
<th>Strand 1</th>
<th>Strand 2</th>
<th>Strand 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Marks</td>
<td>Description</td>
<td>Marks</td>
</tr>
<tr>
<td>6–8</td>
<td>Accurate and detailed knowledge of … is demonstrated.</td>
<td>5–6</td>
<td>The response is highly focused on the question.</td>
</tr>
<tr>
<td>3–5</td>
<td>Generally accurate knowledge of … is demonstrated, with some omissions.</td>
<td>3–4</td>
<td>The response is generally focused on the question.</td>
</tr>
<tr>
<td>1–2</td>
<td>Limited knowledge of … is demonstrated.</td>
<td>1–2</td>
<td>The response relates to the topic but addresses the question only to a limited extent.</td>
</tr>
<tr>
<td>0</td>
<td>No relevant content.</td>
<td>0</td>
<td>No relevant content.</td>
</tr>
</tbody>
</table>

**Figure A2**: Analytic levels-based mark scheme – adapted from the holistic example in Figure A1
Is comparative judgement just a quick form of multiple marking?

Tom Benton Research Division and Tom Gallacher GL Assessment
(The study was completed when the second author was based in the Research Division at Cambridge Assessment)

Introduction

For many years now, comparative judgement (CJ) has been proposed as an alternative method to traditional marking for summative assessment (see, for example, Pollitt, 2004). Traditional marking relies upon an examiner reviewing a piece of work (perhaps an essay) and then making an absolute judgement regarding the mark or grade that it deserves on a pre-determined fixed scale. However, according to proponents of CJ “… humans are very bad at making such absolute judgements” (Christodoulou, 2018). As such, they say it is better to avoid absolute judgements entirely and, instead, assign scores by repeatedly getting examiners to decide which of a pair of scripts (e.g., essays) is superior and then using a statistical model (the Bradley-Terry model) to convert the set of judgements into a single score for each script¹. In particular, this avoids issues with ensuring that all examiners interpret the marking or grade scale in a consistent manner. For example, whereas absolute judgements may be influenced by the leniency or severity of individual examiners, comparative judgements automatically avoid such influences because in CJ, examiners never assign a mark or grade, and make only relative judgements between scripts.

The aim of this article is not to question the claim that CJ can produce valid achievement scores. Rather, the aim is to investigate why it is that CJ works as well as it does, and what would be required for traditional marking to perform equally well. Specifically, this article provides evidence supporting the following central conjecture:

Comparative judgement is just a form of multiple marking, with a very simple mark scheme and using ‘fancy statistics’.

As underlined above, there are three elements to this conjecture. Let us unpack them a little further:

- **Multiple marking:** The crucial term here is ‘multiple’. Although judges in a CJ exercise do not give marks to scripts in the formal sense, multiple judgements are made regarding each piece of work. Usually (although not always) the judgements of several judges are combined to create an overall score for each script.

- **Simple mark scheme:** Rather than having a detailed mark scheme telling judges how to map particular performances to specific scores, judges are generally asked to do something cognitively much simpler – to simply say which of two scripts is superior. This is not to say that CJ exercises do not pay careful attention to defining the skills they are trying to measure. Most CJ exercises include some form of training to help judges know which aspects of performance to focus on. However, at the point of making judgements, the task that judges are required to undertake is clearly simpler than in traditional marking.

- **Fancy statistics:** In traditional marking, the score assigned by a marker is usually exactly the same as the one that is communicated to a candidate. In contrast, judgements from CJ are processed through a fairly complex statistical model to create scores. This model is a crucial part of the machinery for ensuring that the scores ultimately assigned to scripts are all upon a single comparable scale.

The point of the conjecture is that, although the three outlined elements are processed through a fairly complex statistical model to create scores, the same is not true of traditional marks. This is true despite the fact that, in modern on-screen marking systems, scripts are randomly assigned to markers, meaning that it is very simple to build statistical models that would adjust scores accounting for the different ways in which markers have used the mark scale. For example, it would be easy to adjust marks to account for the relative leniency or severity of different markers. To take another example, whereas every CJ exercise requires that multiple judgements are made about each script, in traditional marking each script is usually marked by only one examiner.

Interestingly, studies of the effect of multiple marking on reliability (e.g., The Office of Qualifications and Examinations Regulation [Ofqual], 2014a) tend to be fairly negative about its impact. One possible reason for this negativity, compared with the positive reviews of CJ, is that research on the impact of multiple marking rarely considers the possibility of processing the resulting marks using a statistical model.

This article will show that, if the numbers of judgements for each script and the statistical models from CJ are applied to data from traditional marking, then we can produce scores of equal (or perhaps better) quality without the need for judges to actually make comparative judgements. More specifically, it will show via a pseudo-CJ approach that, if we recreate paired comparisons data such as that used in CJ exercises but based on data from a multiple marking study, we can produce scores that are just as useful as those from a real CJ study. In other words, it is not necessarily the quality of judgements themselves that improve when using CJ; rather it is simply that there are more of them and that statistical models are used to iron out differences in the leniency and severity of different judges. Given that it is in fact the ability to quickly produce judgements for each script that is at the heart of CJ’s strength, the question of whether mark schemes can be simplified to facilitate a greater marking speed naturally arises.

¹ For further technical details of the way the scoring scale is defined, see Bramley and Vitello (2018).
Data

This article makes use of exactly the same set of data as an earlier study by Bramley and Vitello (2018). The data is drawn from 150 essays from a Higher Tier English Language General Certificate of Education (GCSE) paper sat in summer 2014. Candidates were asked to write a diary entry or blog on the subject of “...and it made me change my mind” and the resulting essays were originally marked out of 40. On average the essays were around 500-words long comprising between 1 and 7 pages of writing. The study by Bramley and Vitello explored the relative reliabilities of adaptive comparative judgement (ACJ) and comparative judgements where the pairs being compared were assigned at random (random comparative judgement or RCJ). The two Cj studies were undertaken by separate sets of examiners and scores for all 150 scripts were derived from each method.

Crucially for this research, the same set of scripts was also used in a study by Child, Munro and Benton (2015). The aim of their study was to evaluate the impact of some fairly cosmetic changes to a mark scheme on marking accuracy. As part of this study, all 150 scripts were each marked by 17 examiners who had not been involved in the original marking of this GCSE paper. In addition, all 150 scripts were also marked by the Principal Examiner (PE) for the paper.

As can be seen from our description, for the 150 essays being studied, 4 methods of scoring have been trialled in previous research. Table 1 provides a comparison of the types and number of judges involved in each of the four methods. It also provides details of the number of judgements used to create scores for the two Cj methods. For further details of the design of the Cj studies, see Bramley and Vitello (2018). All the Cj tasks were run online using the No More Marking™ website (www.nomoremarking.com).

Table 1: A comparison of the features of different methods of generating scores

<table>
<thead>
<tr>
<th>Method</th>
<th>Adaptive comparative judgement (ACJ)</th>
<th>Random comparative judgement (RCJ)</th>
<th>Traditional marking</th>
<th>Principal Examiner marking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>On-screen judgements</td>
<td>Paper-based marking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of judges involved in total</td>
<td>18</td>
<td>16</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>Who the judges were</td>
<td>Experienced markers of the given paper</td>
<td>Experienced markers of the given paper</td>
<td>Markers who had never marked the paper in question before</td>
<td>The Principal Examiner</td>
</tr>
<tr>
<td>Average number of judgements on each essay to produce scores for each script in current study</td>
<td>14.4</td>
<td>13.4</td>
<td>Typically 1 (but data also allows analysis of multiple marking)</td>
<td>1</td>
</tr>
</tbody>
</table>

Analyses

Predictive value

The research in this article compares the predictive value of scores derived in different ways (including each of those shown in Table 1). By predictive value, we mean the correlation between the scores assigned to the 150 essays and other external measures of student achievement. Specifically, the focus was upon how well scores derived in each of the above ways correlated with achievements in other tests of English Language and English Literature taken during the same examination session. This measure might be called predictive validity (we are seeing if the scores from the English Language examination can help us ‘predict’ something else that is true about the candidates), although, given that the various assessments being correlated were all taken during the same examination session, others might prefer the term concurrent validity. Certain experts might also take issue with any use of the term validity at all, on the grounds that validity can only be attached to various uses of test scores, rather than the test scores themselves. For this reason, we shall largely avoid the term validity throughout the article and instead use predictive value or correlation to describe the results.

The focus is upon predictive value rather than reliability for two reasons. Firstly, for ACJ, it is almost certain that the reliability coefficients routinely produced by this method are biased and give values that are far too high (Bramley & Vitello, 2018; Bramley, 2015). Secondly, even when restricting our attention to the other methods, we note that the values of reliability coefficients are dependent upon the scale used to report scores. As such, because traditional marking and Cj report scores on different scales, it is not clear that, for example, a scale separation reliability coefficient from Cj can be straightforwardly compared to (say) an inter-marker correlation coefficient from traditional marking. Focussing on the predictive value of scores avoids this issue.

The predictive value of different methods of generating scores [e.g., traditional marking and Cj] was evaluated via their Spearman rank order correlation with:

- the raw marks achieved by each candidate in the controlled assessment element of their English Language GCSE. The controlled assessment task was taken internally within schools, marked by teachers (and then moderated), and tested their skills in understanding extended literary texts and imaginative writing. Scores from this element of the English Language GCSE were available for 128 out of the 150 candidates included in the analysis.
- the Uniform Mark Scale (UMS) marks achieved by candidates across their entire English Literature GCSE. UMS marks summarise achievement across a number of different internal and external assessments within this GCSE and were available for 121 out of the 150 candidates.

Spearman correlations were chosen as they are invariant to the reporting scale used for scores.

Scores from traditional marking

As can be seen, for the Traditional marking column in Table 1, we can choose how many judgements of each essay we combine to produce scores. Several values were trialled for this current research study.
Specifically, scores were produced via:

- **Single marking.** For each script in turn we randomly selected an examiner from the 17 available and used their mark for this script as the score.
- **Double marking.** Each script was assigned a score by taking the mean mark across two randomly chosen examiners.
- **Combining marks from all 17 markers (17-fold marking).** The average mark across all 17 examiners was assigned to each script. Both the median and the mean were trialled.

The use of double marking and 17-fold marking allows us to look at the effect of including one standard feature of CJ (multiple marking) within traditional marking.

In addition, the impact of applying fancy statistics to each of single, double, and 17-fold marking was also trialled. One way of doing this was to standardise each examiner’s marks by subtracting their overall mean and dividing by their standard deviation. The predictive value of scores based upon single and double marking using these scaled marks could then be analysed. Alternatively, statistical processing could be done by treating each marker’s scores as if they were scores from separate items and using a Rasch model to calibrate the ‘difficulty’ of getting each particular mark from each given marker. Then for whichever marker (or combination of markers) was chosen within single, double, or multiple marking, a score could be assigned by combining the mark(s) assigned to the script with these calibrated difficulties. These steps were undertaken using the R package mirt (Chalmers, 2012).

**Pseudo-CJ study**

The central question in this article is whether any improved predictive value of CJ is achieved because of the judgemental process itself (i.e., improving the quality of judgements by making them relative judgements), or whether it is due to the increase in the number of judgements that are made about each script and the way these are analysed. In order to answer this question directly, a final method of creating scores out of traditional marks was trialled – pseudo-CJ. For the purposes of this method, all of the paired comparisons that were completed within the random CJ study were recreated based upon marks. For example, suppose that one row of data within this study indicated that the first judge in the RCJ study has compared Script 131 and Script 20. We then look up the marks assigned to Script 131 and Script 20 by the first marker (not the same individual as the first judge because the judges used in different studies are entirely distinct and are randomly ordered in both cases). Whichever script was awarded the higher mark is treated as if they were judged as superior in a paired comparison study. If both scripts were awarded the same mark by the relevant marker, then the superior script in the pair is chosen completely at random. Suppose that the next row of RCJ data indicates that Judge 10 compared Script 108 and Script 20. We replace this data using information on which of these scripts was awarded the higher mark by Marker 10. This process is continued until the entire data set from the RCJ study has been replaced by paired comparisons based upon marks. The various paired ‘comparisons’ were then converted into scores in the same way as would be done for a real CJ study using the R package sirt (Robitzsch, 2018).

Note that, the RCJ study was designed so that each judge only viewed each script once and, on average, each script was judged by 13.4 of the judges. As such, in the pseudo-CJ data, on average each script has data from paired comparisons of their marks against 13.4 other scripts based on the marking from 13.4 markers. It is very important for this element of the study that it is based on genuine instances of multiple marking. While it would be possible to convert a single marker’s marking into a set of paired comparisons, it would be pointless and would ultimately result in the same rank order of candidates as the original marks. For this method, it is the fact that the pseudo-CJ method combines the judgements of multiple markers that leads to the expectation that it will have increased predictive value relative to single marking.

The amount of examiner time required to produce scores is an important consideration in any research into different methods of marking. After all, regardless of how reliable it may be, there is little point in suggesting a method that would require hours of examiner time for every single essay. So, in order to provide context for the analyses we have described, we also examined the amount of time taken to complete traditional marking for each script and how this compared to the length of time required to complete each individual paired comparison.

In addition, the analysis calculated the correlation between CJ scores and traditional marks in order to investigate whether the two approaches appeared to be generally measuring the same skills. Finally, in order to test whether scores from CJ might be more strongly influenced by superficial features of essays, the analysis investigated whether the association between essay length and scores differed depending upon whether CJ or traditional marking was used.

**Results**

**Speed of marking**

To begin with, since it affects our interpretation of results regarding predictive value, we present some information regarding the speeds of CJ, and of traditional marking. As part of the software used to undertake the CJ exercises, the time taken to make each individual judgement was recorded. The mean amount of time for each judgement was recorded as 3.5 minutes. Given that, as mentioned earlier, the essays were each around 500-words long (so each pair of essays consisted of about 1,000 words), this indicates that judgements took about 1 minute for every 300 words of written text. According to results published at http://www.readingsoft.com/, this is close to a typical on-screen reading speed for a good reader.

As part of the study by Child et al. (2015), examiners were asked to complete a questionnaire which included a question on how long they felt it took them to mark each script. It is important to note that markers were not only asked to mark the essays being studied in this article (Question 4 in the original exam paper) but also another slightly shorter question (Question 2). The results in Table 2 reflect the amount of time examiners felt that it took them to mark both of these questions. As can

---

2. A more complex approach based upon the graded response model (GRM) was also trialled. However, it was not found to have any noticeable impact upon results and so, for brevity, the results from this method are omitted from this article.

3. A fairly recent blog has suggested that paired comparisons of GCSE English essays can be done within as little as 23 seconds on average (https://blog.nomosmarking.com/judging-gcse-english-efficiency-and-reliability-9a4df19f80d098/). With essays of the length of those in our research, this would mean judges were reading at almost 3,000 words per minute – a speed that is not credible (equivalent to reading this entire footnote in 2.2 seconds). It is possible that the research in the blog is based upon much shorter essays than our research. Alternatively, it may be that, in contrast to our research, judges saw the same essay many times and so could work from memory rather than re-reading each essay in full every time.
be seen, most examiners responded either that it took “6–10 minutes” or that it took “11–20 minutes” to mark both of these questions. Given that there are slightly more respondents in the “11–20” category, we might estimate that it took markers around 11 minutes to mark each script on average. Table 2 gives information on the amount of time it took markers to mark two questions in one script: Question 2 (an essay marked out of 14) and Question 4 (an essay marked out of 40). If we assume that the amount of time taken to mark each of these questions is roughly proportional to the available number of marks, we can derive a rough estimate that it took about 7 minutes on average for an individual examiner to complete traditional marking for one essay.

Table 2: Examiners’ responses regarding how long it took them to complete traditional marking

<table>
<thead>
<tr>
<th>Typically, how long did it take you to mark one script (Questions 2 and 4)?</th>
<th>Number of examiners</th>
</tr>
</thead>
<tbody>
<tr>
<td>3–5 minutes</td>
<td>2</td>
</tr>
<tr>
<td>6–10 minutes</td>
<td>6</td>
</tr>
<tr>
<td>11–20 minutes</td>
<td>8</td>
</tr>
<tr>
<td>Over 20 minutes</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
</tr>
</tbody>
</table>

Recall that each paired comparison in the CJ exercise took 3.5 minutes to complete, whereas we estimated that traditional marking for a single one of these essays took about 7 minutes (i.e., twice as long). Also note that each paired comparison provides a judgement for two essays, whereas marking only deals with one essay at a time. Putting these facts together indicates that CJ produces judgements for essays at about four times the speed of traditional marking.

Predictive value

The Spearman correlations of each method of producing scores for the English essays with scores on the two external achievement variables are shown in Table 3. To enable some of the main patterns to be seen more clearly, the same information is presented visually in Figure 1.

To begin with, we note that the predictive value of RCJ was higher than that from ACJ for both the English Language and English Literature assessments. This was despite ACJ having a reported (scale separation) reliability coefficient of 0.97 compared to 0.72 for the RCJ method. This confirms the conclusions from Bramley and Vitello (2018) that reliability coefficients from ACJ are biased upwards. It is simply not credible that ACJ could display improved reliability to this extent without it translating into any meaningful improvement in predictive value.

Before going too much further, it is worth noting that, as would be hoped by proponents of CJ, both the ACJ and RCJ methods provided greater predictive value than the raw mark values from traditional single marking. The difference between the predictive value of CJ and that of traditional marking was similar in both direction and scale to that reported by Steedle and Ferrara (2016). However, it is notable that all of the advantage of ACJ and almost half of the advantage of RCJ vanished once these marks were scaled either using the simple mean/SD method, or by using Rasch analysis. In other words, much of the apparent advantage of CJ can be explained by its use of fancy statistics. Just as statistical methods are used in CJ to ensure that scores are on a consistent scale, statistical methods can also be applied to traditional marks to ensure the same thing and, at least in this data set, doing this improved the predictive value of the marks.

Perhaps most importantly for our research question, we see that the pseudo-CJ method based upon marking yielded predictive values similar to the RCJ method. In other words, if the number of judgements for each essay and the method of analysis are held constant, it makes no difference whether the data comes from actual paired comparisons or from pseudo-comparisons derived from marks. This implies that there is nothing magical about placing two essays next to each other that allows humans to make better decisions regarding their quality. Rather, the benefit comes from the number of judgements that are made and the ways these are combined. Of course, it may well be that it is the simplicity of the paired comparison task, and the resultant increased speed of judgements, that facilitates collating this number of judgements.

The predictive value of RCJ is similar to that of double marking if marks are combined without any form of scaling. If double marking is combined with any form of scaling, its predictive value increases further. Table 3 also shows estimates of the amount of time taken by judges to produce the scores given by each method. As noted earlier, we expect that each marker took around 7 minutes to mark each essay meaning that double marking would take 14 minutes for each essay. We have also seen that each paired comparison took 3.5 minutes (or 1.75 minutes for each essay being compared) so that we would expect either CJ task to require more than 20 minutes of examiner time for each script on average. In other words, the time expended on each CJ method was roughly equivalent to the time required to complete triple marking. However, the predictive value of the method was only equivalent to that of double marking.

The predictive value of combining the marks from all markers was even greater. This is unsurprising as we would expect using a larger number of markers to improve the marking reliability of the resulting scores. Finally we note that, although the predictive value of marks from the PE was above that of double marking, the predictive value of their marks was not as high as that of the consensus mark derived across all markers.

Comparison of rank order of CJ score with combined mark from all markers

As noted in Table 3, the predictive value of RCJ scores was lower than that of marks derived from all markers combined – for example, using the mean of all awarded marks. With this in mind, it is of interest to explore whether this is because RCJ measures something fundamentally different to marking (perhaps rewarding different script features), or whether this can just be explained in terms of the relative reliability of the two methods. This question is explored in Figure 2 which shows the association between the two sets of scores. As can be seen, there was a relatively strong correlation between the two measures (0.80). To interpret this correlation, we note that the reported reliability of the RCJ method was 0.721. Thus, if the method of taking the mean of the 17 marks was perfectly reliable, and the two methods were measuring precisely the same thing, we would expect a correlation of \[
\sqrt{0.721} = 0.85
\] between the two methods. That the actual correlation is not far off this value suggests that the RCJ method measures essentially the same skills as those rewarded in traditional marking.
Table 3: Different methods of scoring the essays and the Spearman correlations of the scores with marks achieved in other tests

<table>
<thead>
<tr>
<th>Scoring method for English Language examination essays</th>
<th>Average number of judgments on each essay</th>
<th>Estimated average time spent on each essay (minutes)</th>
<th>Correlation of score with...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>English Language Controlled Assessment (N=121)</td>
</tr>
<tr>
<td>Comparative judgement</td>
<td>ACJ</td>
<td>14.4</td>
<td>25.2</td>
</tr>
<tr>
<td></td>
<td>RCJ</td>
<td>13.4</td>
<td>23.5</td>
</tr>
<tr>
<td>Pseudo-CJ based on marks</td>
<td></td>
<td>13.4</td>
<td>93.8</td>
</tr>
<tr>
<td>Single marking</td>
<td>Raw marks</td>
<td>1</td>
<td>7.0</td>
</tr>
<tr>
<td></td>
<td>Scaled marks (mean/SD)</td>
<td>1</td>
<td>7.0</td>
</tr>
<tr>
<td></td>
<td>Rasch scaled marks</td>
<td>1</td>
<td>7.0</td>
</tr>
<tr>
<td>Double marking</td>
<td>Raw marks</td>
<td>2</td>
<td>14.0</td>
</tr>
<tr>
<td></td>
<td>Scaled marks (mean/SD)</td>
<td>2</td>
<td>14.0</td>
</tr>
<tr>
<td></td>
<td>Rasch scaled marks</td>
<td>2</td>
<td>14.0</td>
</tr>
<tr>
<td>All markers</td>
<td>Rasch model</td>
<td>17</td>
<td>119.0</td>
</tr>
<tr>
<td></td>
<td>Median mark</td>
<td>17</td>
<td>119.0</td>
</tr>
<tr>
<td></td>
<td>Mean mark</td>
<td>17</td>
<td>119.0</td>
</tr>
<tr>
<td>Principal Examiner’s marks</td>
<td></td>
<td>1</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

Figure 1: Spearman correlations of the scoring methods with external achievement variables

Figure 2: The association between scores from RCJ and the mean mark awarded to candidates across all 17 markers (i.e., not including the PE) (Note: Spearman correlation=0.80).
Relationship with length of responses

One concern over the use of CJ is that it is harder to check whether the features of responses that are used to make ratings match with those that are intended. For example, in the context of a study of comparative judgements of videos of an English speaking assessment, Pollitt and Murray (1996) mentioned that “it was no surprise to find evidence that the judges were influenced to some extent by the candidates’ personalities, physical attractiveness, nationalities, and cultural backgrounds”. In the case of our research, we might be concerned (as other authors have been) that judges taking part in a CJ exercise may be particularly susceptible to influence from extraneous factors, such as handwriting or the sheer length of responses (Bramley, 2007).

Although no measure of handwriting quality was available, one simple check that could be performed was whether the association of essay length with essay scores was consistent across the different methods of rating them. In particular, if judges within the CJ exercise were using essay length as a shortcut to make judgements, rather than fully considering the extent to which candidates have displayed the desired set of skills, then we might see a stronger association between essay lengths and scores from CJ than is the case with traditional marking. Of course, given the results shown in Figure 2, we have already stated that it appears that both CJ and marking are measuring the same set of skills. However, exploring the relationship with essay length provides another check.

The approximate word count of each essay was calculated using a method similar to the one described in Benton (2017). Figure 3 shows a scatter plot of these approximate word counts against standardised scores from three methods: RCJ, the mean mark from all markers (except the PE), and the PE’s marks. Scores were standardised by subtracting the mean and dividing by the standardisation. Standardised scores are used in Figure 3 as it puts results from all of the different scoring methods on the same scale.

The relationship between essay length and scores in Figure 3 is reminiscent of a similar chart shown in Benton (2017) for English Literature essays. As was seen in the earlier research, very short essays tended to be awarded lower marks than those of average length or longer. However, there is no evidence that very long essays were awarded higher scores than those of average length. More importantly for our research, we can see that the link between essay length and scores was very similar regardless of which method was used to generate them. In particular, there is no evidence of CJ being any more likely to favour long essays than the PE. This supports the earlier evidence that, in this study, both CJ and marking were measuring the same set of skills. Given that the judges involved in the CJ exercise were already experienced markers of the examination being studied, perhaps this similarity should come as no surprise.

Conclusion

The central conjecture of this article was that CJ is just a form of multiple marking combined with a simple mark scheme and fancy statistics. The evidence from this study supports this conjecture. In particular, we have seen that if we recreate the paired comparisons in a CJ study using marks from a multiple marking study, the resulting scores from such a pseudo-CJ exercise have at least as much predictive value as scores from an actual CJ exercise. In other words, it is the number of judgements that are made about scripts and the way they are analysed within a CJ exercise that is important. The physical act of placing two essays next to each other and deciding which is better does not appear to produce judgements that, in themselves, have any more predictive value than getting the same individual to simply mark a set of essays.

Having seen that the strength of CJ lies in its use of multiple judgements and a strong statistical model, the evidence in this article suggests that these same techniques could be employed in traditional marking to improve the reliability and validity of scores. This need not be achieved by fully recreating CJ studies. For example, a statistical model can be applied to scores from traditional marking through much simpler techniques, such as marker scaling to address differences in the leniency and severity of different examiners.

The key question in whether CJ could replace marking is which of the two methods makes better use of the resources available to us. In this particular study, it appears that multiple marking was the more efficient. In particular, whilst the various CJ exercises took roughly as long as triple marking, they only achieved the predictive value of double marking. Thus, in this case, double marking (combined with statistical scaling) would appear to be the more efficient method.

It is important to note that, this is only a single study and has only considered CJ and marking for a particular task – marking GCSE English essays. Alterations to the design of either the CJ study or the marking study may lead to different results. For example, what would happen if judges in the CJ study were explicitly encouraged to make decisions more quickly? Could this be done without harming the reliability and validity of the resulting scores? Or in the case of marking, what if the mark scheme was simplified to encourage marking to be done more quickly with the express intention of subsequently using statistical scaling to iron out differences in leniency and severity between examiners? Indeed, given that our evidence shows that, provided they are analysed properly, absolute judgements can be just as useful as relative judgements, it is natural to ask how the process of producing absolute judgements can be...
made quicker. For example, would getting examiners to give an intuitive score for each essay on a scale from 1–10, and then processing these scores using a Rasch model result in an equally useful set of student scores as traditional marking, but in a fraction of the time?

It is, of course, not necessarily true that the results shown here with respect to GCSE English essays would be repeated for other subjects. It is doubtful the case that certain types of student performances lend themselves more readily to CJ, whereas others are easier to mark. Nonetheless, the results here are important in understanding where the benefits of CJ derive from. Recognising that these are not solely caused by switching the way in which judgements are elicited, but also in the number of such judgements and how they are analysed, allows for a more nuanced comparison of the relative advantages of CJ and traditional marking. Failing to recognise these differences risks the two approaches never being compared on a like-for-like basis.

This article should not be taken as a criticism of the existing system and Curriculum Authority.

References


How have students and schools performed on the Progress 8 performance measure?

Tim Gill  Research Division

Introduction

In October 2013, the Department for Education (DfE) announced that new ‘headline’ performance measures for schools would be introduced to replace the previous measure of the percentage of students achieving five or more grades A* to C at General Certificate of Secondary Education (GCSE) Level, including English and Mathematics. The new measures (known as Attainment 8 and Progress 8) are based on performance in a student’s best eight subjects at GCSE (or equivalent), although with some restrictions, students are required to take the English Baccalaureate (EBacc) qualifications in English and Mathematics, as well as at least three other EBacc qualifications. The remaining three slots can be filled either by other EBacc qualifications or by other approved, non-EBacc qualifications.

One of the reasons for the introduction of the new measures was concern that the previous measures penalised schools with a low-attaining intake. As Progress 8 is a value-added measure, it already accounts for the prior attainment of the student and should, in theory, no longer penalise these schools. The following method is used to calculate school-level Progress 8 scores:

- Calculate the Attainment 8 score for each student. This is the total points score for their eight highest scoring eligible qualifications. Points are based on the grade achieved (e.g., for GCSEs, points are on a 1–8 scale; 1 = G, 8 = A*2).

1. EBacc is the English Baccalaureate, a school performance measure which shows the proportion of pupils studying the ‘core’ academic subjects at KS4. Only specific qualifications (mainly GCSEs) are eligible for inclusion in the EBacc.
2. This example is for ‘old’ GCSEs. The scores for new GCSEs (9–1 grading scale) are slightly different.


- Compare this score with the mean Progress 8 score for students with the same prior attainment (as measured by Average Fine Level on Key Stage 2 [KS2] tests). The Progress 8 score for a student is the difference between the two, divided by 10 (this turns the score into a per qualification measure, as the English and Mathematics points are double-weighted in the calculation). This division is always by 10, even if a student takes fewer than 8 eligible qualifications.

- Calculate the mean Progress 8 score for all students in the school. This is the school’s Progress 8 score.

By definition, the mean Progress 8 score for students with the same prior attainment is always zero. However, despite the value-added structure of Progress 8, there have been various criticisms levelled at the new measure. At the level of the school, there is evidence that it is biased towards Selective schools (e.g., Allen, 2016; Andrews, 2017). At student level, there is evidence that certain groups perform systematically better than others. These include female students, and those of Chinese ethnicity (Andrews, 2017), non-free school meals (non-FSM) students (Andrews, 2017; Sherrington, 2017), and non-Pupil Premium and EAL students (Thomson, 2017).

A further criticism of Progress 8 is that a few student-level outliers can have undue influence on the school-level score. Both Allen (2017) and Sherrington (2017) found that having just a handful of students who for one reason or another did not sit any eligible qualifications (and therefore achieved Attainment 8 scores of zero) can severely reduce a school’s overall Progress 8 score. A recent UK Government policy document on school performance tables (DfE, 2017a) revealed that the DfE plans to consult with schools on this issue with a view to making changes to the methodology in future years.

The purpose of this research was to delve deeper into the relationship between Progress 8 scores and various student- and school-level factors. Prior research tended to focus on basic differences in Progress 8 mean scores between groups. This article presents similar analyses on more recent data and builds on this with a more detailed analysis, including a linear regression model to infer which factors were most important in determining scores at student level.

Data and methods

We used data from the 2015/16 academic year. We were interested in both the school- and student-level Progress 8 scores. Students’ Progress 8 scores were taken directly from the National Pupil Database (NPD), which is administered by the DfE. The NPD includes examination results for all students in all qualifications and subjects in schools and colleges in England, as well as student and school background characteristics such as gender, ethnicity and level of income-related deprivation.

Data on school-level Progress 8 scores was downloaded from a DfE website (https://www.compare-school-performance.service.gov.uk/download-data).

We used school classification information from Edubase (the DfE’s register of educational establishments), which classifies schools by their school ‘type’ and by their selection policy. We excluded data on Independent schools and Special schools because they are not subject to the same accountability measures as State schools. We also excluded data on further education (FE) and sixth form colleges because, although some offer GCSEs or other qualifications to students in Key Stage 4 (KS4), this is a relatively rare occurrence.

The geographical region of each school was downloaded from a UK Government website (https://get-information-schools.service.gov.uk). We undertook a descriptive analysis of Progress 8 scores for different groups of students and schools, followed by statistical modelling of Progress 8 scores at student level. We categorised students by a number of different background characteristics recorded in the NPD:

- **FSM status:** We classified students by whether or not they had claimed for free school meals (FSM) in any of the past six years.

- **SEN status:** We classified students with Special Educational Needs (SEN) by the categories used in the NPD. These were (in order of the amount of extra support needed, from low to high): SEN support; Statement of SEN; or Education, Health and Care Plan (EHCP).

- **Ethnicity:** The NPD categorised students into one of seven ethnic groups: White; Asian; Black; Chinese; Mixed; Other; or Unknown. Chinese students were in a category of their own due to a well-known tendency to perform very well compared to other Asian students.

- **EAL status:** The English as an Additional Language (EAL) classification in the NPD was into one of three categories: English; Other; or Unclassified. It should be noted (see Strand, Malmberg & Hall, 2015) that the definition of EAL in the NPD only accounted for whether the student was exposed to an additional language in their home or community. It did not actually tell us their level of proficiency in English.

- **School type:** We classified students by their school type, taken from Edubase: Academy (Comprehensive); Academy (Modern); Academy (Selective); Comprehensive; Secondary Modern; and Grammar.

- **School gender:** Girls’; Boys’; or Mixed.

- **Region:** South East; London; North West; East of England; West Midlands; South West; Yorkshire and the Humber; East Midlands, or North East.

Additionally, we investigated the relationship between two non-categorical variables and Progress 8 scores. These were:

- **Prior attainment:** Measured by students’ KS2 Average Fine Level. This was not included in the descriptive analysis because it is already accounted for in the calculation of Progress 8, and therefore average Progress 8 scores do not vary with levels of prior attainment.

- **Income-related deprivation:** A measure of deprivation commonly used in analyses of student performance is the Income Deprivation Affecting Children Index (IDACI). This measures the percentage of children in the area where the student resides who live in income-deprived families. As such, it cannot tell whether or not the students themselves are income deprived.

For the school-level analysis, we calculated various measures to indicate the composition of students in each school, as the types of students attending a school were likely to have a significant impact on the Progress 8 score for that school. They may also have an impact on the Progress 8 scores for individual students. We calculated the following school level variables (using the data in the NPD on students at the end of KS4):

3. Average Fine Levels are derived from the marks achieved on KS2 tests in Mathematics and English.

4. For further information on IDACI calculation, including definitions of children, families, and income deprivation, see https://www.gov.uk/government/publications/english-indices-of-deprivation-2015-technical-report
described above we remost important in predicting the student-level performance on Progress 8. A multilevel model was used to account for the clustering of students within schools. The general form of the model was as follows:

\[ y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + \mu_j + \epsilon_{ij} \]

Where \( y_{ij} \) is the Progress 8 score for student \( i \) in school \( j \), \( x_{1ij}, \ldots, x_{kij} \) are the independent variables, \( \beta_1, \ldots, \beta_k \) are the regression coefficients, \( \mu_j \) is a school effect (this is technically known as the Level 2 random effect and allows the model to account for the clustering of pupils within schools) and \( \epsilon_{ij} \) is the residual difference between a student’s predicted and actual Progress 8 score.

Results

Descriptive analysis of students’ performance on Progress 8

This section presents the results of descriptive analyses of student-level Progress 8 scores, by various background factors. This summary is similar to that provided by the DfE (see https://www.gov.uk/government/statistics/revised-gcse-and-equivalent-results-in-england-2015-to-2016), although their analysis included data from Special schools and FE colleges, while ours does not.

Table 1 presents the number of students in each category, and the mean and standard deviation (SD) of Progress 8 (P8) scores, by each of the student-level factors. This shows a small, but noteworthy, difference between the mean scores for girls (0.13) and boys (-0.12). This means that on average, girls made the equivalent of a quarter of a grade more progress per subject than boys.

Around a quarter of students were in the FSM category and, on average, they had substantially lower Progress 8 scores (-0.31) than students in the non-FSM category (0.12).

There were around 13% of students categorised as having some level of SEN, of which most were in the SEN support category (11%). Only 1.6% of students had a Statement of SEN and 0.4% had an EHCP. It is clear that students with no SEN had a much higher mean Progress 8 score than any of the SEN students. Students with EHCP had the lowest mean Progress 8 score, followed by those with SEN support and those with a Statement.

Table 1: Distribution of Progress 8 scores, by student characteristics

<table>
<thead>
<tr>
<th>Category</th>
<th>No. of students</th>
<th>Mean P8</th>
<th>SD P8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>248,547</td>
<td>0.13</td>
<td>1.02</td>
</tr>
<tr>
<td>Male</td>
<td>253,257</td>
<td>-0.12</td>
<td>1.08</td>
</tr>
<tr>
<td><strong>FSM</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>367,843</td>
<td>0.12</td>
<td>0.95</td>
</tr>
<tr>
<td>Yes</td>
<td>133,406</td>
<td>-0.31</td>
<td>1.24</td>
</tr>
<tr>
<td><strong>SEN status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>435,912</td>
<td>0.06</td>
<td>1.00</td>
</tr>
<tr>
<td>SEN support</td>
<td>55,397</td>
<td>-0.38</td>
<td>1.32</td>
</tr>
<tr>
<td>Statement</td>
<td>8,019</td>
<td>-0.31</td>
<td>1.31</td>
</tr>
<tr>
<td>EHCP</td>
<td>2,044</td>
<td>-0.49</td>
<td>1.40</td>
</tr>
<tr>
<td><strong>Ethnic group</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>45,726</td>
<td>0.34</td>
<td>0.98</td>
</tr>
<tr>
<td>Black</td>
<td>23,688</td>
<td>0.21</td>
<td>1.04</td>
</tr>
<tr>
<td>Chinese</td>
<td>1,584</td>
<td>0.70</td>
<td>0.83</td>
</tr>
<tr>
<td>Mixed</td>
<td>20,858</td>
<td>-0.01</td>
<td>1.12</td>
</tr>
<tr>
<td>White</td>
<td>399,719</td>
<td>-0.06</td>
<td>1.05</td>
</tr>
<tr>
<td>Other</td>
<td>6,196</td>
<td>0.50</td>
<td>1.06</td>
</tr>
<tr>
<td>Unknown</td>
<td>4,236</td>
<td>-0.05</td>
<td>1.16</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>436,739</td>
<td>-0.06</td>
<td>1.05</td>
</tr>
<tr>
<td>Other</td>
<td>63,842</td>
<td>0.42</td>
<td>1.03</td>
</tr>
<tr>
<td>Unclassified</td>
<td>862</td>
<td>0.02</td>
<td>1.18</td>
</tr>
<tr>
<td><strong>School type</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Academy (Comprehensive)</td>
<td>290,343</td>
<td>0.01</td>
<td>1.06</td>
</tr>
<tr>
<td>Academy (Selective)</td>
<td>17,467</td>
<td>0.33</td>
<td>0.72</td>
</tr>
<tr>
<td>Academy (Modern)</td>
<td>12,251</td>
<td>-0.03</td>
<td>1.06</td>
</tr>
<tr>
<td>Comprehensive</td>
<td>172,900</td>
<td>-0.04</td>
<td>1.08</td>
</tr>
<tr>
<td>Grammar</td>
<td>3,005</td>
<td>0.32</td>
<td>0.75</td>
</tr>
<tr>
<td>Secondary Modern</td>
<td>5,838</td>
<td>-0.14</td>
<td>1.10</td>
</tr>
<tr>
<td><strong>School gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>451,353</td>
<td>-0.03</td>
<td>1.07</td>
</tr>
<tr>
<td>Boys</td>
<td>20,096</td>
<td>0.15</td>
<td>0.96</td>
</tr>
<tr>
<td>Girls</td>
<td>30,355</td>
<td>0.31</td>
<td>0.95</td>
</tr>
<tr>
<td><strong>Region</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>East Midlands</td>
<td>44,612</td>
<td>-0.11</td>
<td>1.06</td>
</tr>
<tr>
<td>East of England</td>
<td>57,473</td>
<td>0.05</td>
<td>1.03</td>
</tr>
<tr>
<td>London</td>
<td>68,467</td>
<td>0.20</td>
<td>1.08</td>
</tr>
<tr>
<td>North East</td>
<td>24,552</td>
<td>-0.11</td>
<td>1.04</td>
</tr>
<tr>
<td>North West</td>
<td>69,862</td>
<td>-0.12</td>
<td>1.08</td>
</tr>
<tr>
<td>South East</td>
<td>79,486</td>
<td>0.05</td>
<td>1.04</td>
</tr>
<tr>
<td>South West</td>
<td>49,264</td>
<td>-0.02</td>
<td>1.04</td>
</tr>
<tr>
<td>West Midlands</td>
<td>56,281</td>
<td>-0.04</td>
<td>1.03</td>
</tr>
<tr>
<td>Yorkshire and the Humber</td>
<td>51,753</td>
<td>-0.01</td>
<td>1.07</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>501,804</td>
<td>0.00</td>
<td>1.06</td>
</tr>
</tbody>
</table>

5 Totals by characteristic may not always add up to the total number of students included, due to missing data.
It is interesting that students with a Statement (i.e., a higher level of special needs) made better progress on average than those receiving SEN support (lower level).

White students made up about 80% of the population, with 9% Asian and around 5% Black. Only about 0.3% of students were Chinese. The results show that Chinese students had the highest scores on average, followed by those in the ‘Other’ category and Asian students. The lowest mean was for White students (-0.12). Thus, all other ethnicities made more progress than Whites. However, this analysis takes no account of other factors, such as deprivation levels, which may be more important in determining Progress 8 scores.

The results in Table 1 show that around 87% of students were English speakers. Non-English speakers had a higher mean (0.42) than English speakers (-0.06).

In terms of school gender, students in both types of single-sex schools had higher Progress 8 scores on average than those in mixed schools. This is likely to be due to a high proportion of single-sex schools also being selective, and students in these schools made more progress on average (see Table 1). Students in girls’ schools made more progress than those in boys’ schools, which may be partly because girls made more progress than boys (see Table 1).

Finally, the differences in mean Progress 8 scores between regions were generally quite small. However, students in London had a notably higher mean (0.20) than any other region. The regions with the lowest means were North West (-0.12), East Midlands, and North East (both -0.11). This analysis takes no account of the background characteristics of the students in each region, which may be more important in determining Progress 8 scores.

Figure 1 displays a plot of the relationship between the implied deprivation experienced by students (as measured by IDACI) and their Progress 8 score. This shows a small, negative relationship between the two measures (correlation = -0.14), indicating that more deprived students had lower Progress 8 scores on average.

![Figure 1: Relationship between student IDACI and Progress 8 score](image)

**Descriptive analysis of schools’ performance on Progress 8**

Table 2 presents descriptive analyses of Progress 8 (P8) scores by various school-level factors (school type, school gender, and school region).

The mean scores by school type were very similar to the mean scores in the student-level analysis (see Table 1). Schools classified as Selective academies or Grammar schools had higher mean Progress 8 scores on average (0.33, compared with negative mean scores for all other school types). The lowest mean was for non-academy Secondary Modern schools (-0.14). There were very few Grammar schools or Selective academies with Progress 8 scores below zero. In contrast, the maximum Progress 8 scores for Secondary Modern schools was just 0.59.

In terms of school gender, girls’ schools had the highest mean Progress 8 score, followed by boys’ schools. The difference between the average Progress 8 score in girls’ schools and in mixed schools (0.36) was equivalent to more than one third of a grade. These results were very similar to the results for students attending single-sex or mixed schools (see Table 1), but it is worth noting that there were very few girls’ schools with a negative Progress 8 score.

Schools in London had the highest mean Progress 8 scores by some distance (0.19). The only other regions with a positive mean were South East (0.04) and East of England (0.02). The regions with the lowest mean were North West (-0.15), East Midlands, and North East (both -0.14). The ‘London effect’ is a well-researched phenomenon that has been attributed to a number of different factors (see Blandon, Greaves, Gregg, MacMillan & Sibeta, 2015), and the results here suggest it is present in terms of progression as well as attainment.

The remaining school-level factors that we investigated were continuous variables, which are better analysed through correlation coefficients and scatter plots. Table 3 presents the correlation coefficients between Progress 8 scores and the value of each of these factors.

### Table 2: Distribution of mean Progress 8 scores, by school characteristics

<table>
<thead>
<tr>
<th>School type</th>
<th>No. of schools</th>
<th>Mean P8</th>
<th>SD P8</th>
<th>Min P8</th>
<th>Max P8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy</td>
<td>1,745</td>
<td>-0.03</td>
<td>0.41</td>
<td>-2.51</td>
<td>1.37</td>
</tr>
<tr>
<td>(Comprehensive)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Selective)</td>
<td>139</td>
<td>0.33</td>
<td>0.20</td>
<td>-0.24</td>
<td>0.75</td>
</tr>
<tr>
<td>Academy (Modern)</td>
<td>79</td>
<td>-0.03</td>
<td>0.32</td>
<td>-1.00</td>
<td>0.59</td>
</tr>
<tr>
<td>Comprehensive</td>
<td>1,053</td>
<td>-0.05</td>
<td>0.34</td>
<td>-1.36</td>
<td>1.08</td>
</tr>
<tr>
<td>Grammar</td>
<td>23</td>
<td>0.33</td>
<td>0.21</td>
<td>-0.06</td>
<td>0.81</td>
</tr>
<tr>
<td>Secondary Modern</td>
<td>43</td>
<td>-0.14</td>
<td>0.35</td>
<td>-0.83</td>
<td>0.54</td>
</tr>
<tr>
<td>Mixed</td>
<td>2,722</td>
<td>0.05</td>
<td>0.39</td>
<td>-2.51</td>
<td>1.31</td>
</tr>
<tr>
<td>Boys’</td>
<td>151</td>
<td>0.16</td>
<td>0.32</td>
<td>-0.60</td>
<td>1.15</td>
</tr>
<tr>
<td>Girls’</td>
<td>209</td>
<td>0.31</td>
<td>0.26</td>
<td>-0.49</td>
<td>1.37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>No. of schools</th>
<th>Mean P8</th>
<th>SD P8</th>
<th>Min P8</th>
<th>Max P8</th>
</tr>
</thead>
<tbody>
<tr>
<td>East Midlands</td>
<td>268</td>
<td>-0.14</td>
<td>0.37</td>
<td>-1.23</td>
<td>0.93</td>
</tr>
<tr>
<td>East of England</td>
<td>345</td>
<td>0.02</td>
<td>0.41</td>
<td>-2.32</td>
<td>0.75</td>
</tr>
<tr>
<td>London</td>
<td>429</td>
<td>0.19</td>
<td>0.35</td>
<td>-1.71</td>
<td>1.4</td>
</tr>
<tr>
<td>North East</td>
<td>147</td>
<td>-0.14</td>
<td>0.35</td>
<td>-1.08</td>
<td>0.69</td>
</tr>
<tr>
<td>North West</td>
<td>445</td>
<td>-0.15</td>
<td>0.42</td>
<td>-2.51</td>
<td>1.37</td>
</tr>
<tr>
<td>South East</td>
<td>473</td>
<td>0.04</td>
<td>0.34</td>
<td>-1.14</td>
<td>1.08</td>
</tr>
<tr>
<td>South West</td>
<td>307</td>
<td>-0.04</td>
<td>0.33</td>
<td>-1.19</td>
<td>0.91</td>
</tr>
<tr>
<td>West Midlands</td>
<td>371</td>
<td>-0.05</td>
<td>0.37</td>
<td>-1.79</td>
<td>1.31</td>
</tr>
<tr>
<td>Yorkshire and the Humber</td>
<td>295</td>
<td>-0.02</td>
<td>0.36</td>
<td>-2.09</td>
<td>0.93</td>
</tr>
<tr>
<td>All</td>
<td>3,082</td>
<td>-0.02</td>
<td>0.38</td>
<td>-2.51</td>
<td>1.37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>School factor</th>
<th>Correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean prior attainment</td>
<td>0.33</td>
</tr>
<tr>
<td>FSM students (%)</td>
<td>-0.28</td>
</tr>
<tr>
<td>SEN students (%)</td>
<td>-0.21</td>
</tr>
<tr>
<td>EAL students (%)</td>
<td>0.23</td>
</tr>
<tr>
<td>Mean IDACI</td>
<td>-0.27</td>
</tr>
<tr>
<td>White students (%)</td>
<td>-0.23</td>
</tr>
<tr>
<td>Asian students (%)</td>
<td>0.18</td>
</tr>
<tr>
<td>Chinese students (%)</td>
<td>0.17</td>
</tr>
<tr>
<td>Black students (%)</td>
<td>0.14</td>
</tr>
</tbody>
</table>
variables at school level. Figure 2 plots the data for four of these factors (prior attainment, percentage FSM, percentage SEN, and percentage EAL).

To give an idea of what the prior attainment scale means, a student who only just achieved Level 4 in English and Mathematics (the minimum expected level in each subject) would have an Average Fine Level equal to 4.0. Figure 2 shows a distinct positive relationship between the two, suggesting that schools with higher prior attaining students also tended to get higher Progress 8 scores on average.

The correlation between the two measures was 0.33. The slope of the line of best fit was 0.48, meaning that an increase in KS2 mean of 1 led to (on average) an increase in Progress 8 score of 0.48. Note that there appeared to be two separate groups of schools, with the main body of schools having a KS2 mean score of between 4 and 5, and a smaller group with a KS2 mean score of between 5 and 5.5. Almost all of this second group were Selective schools, which were able to select exclusively high-attaining students.

There was a clear negative relationship between the school percentage of FSM students and Progress 8 scores (correlation = -0.28). However, we can see from Figure 2 that there were many schools with high percentages of FSM students and positive Progress 8 scores, and many with low percentages of FSM students and negative Progress 8 scores.

There was also a significant negative relationship between the percentage of students with SEN in a school and Progress 8 score, with a correlation of -0.21. In other words, schools with higher percentages of SEN students tended to have lower Progress 8 scores. This is not a surprising finding, as students with special needs made less progress on average than other students (see Table 1).

There was a significant positive relationship between the percentage of EAL students and Progress 8 scores (correlation = 0.23). Schools with higher percentages of EAL students were more likely to have higher Progress 8 scores than those with low percentages of EAL. We expected this result, given that EAL students made considerably more progress on average than students with English as their first language (see Table 1).

Table 3 also shows that the correlation between the mean level of implied deprivation (as measured by IDACI) experienced by students attending the school and the school Progress 8 score was -0.27. Given the negative relationship between disadvantage and Progress 8 score at an individual level, it is not surprising that the same relationship was also visible at school level.

The ethnic make-up of schools was also considered as a possible influence on Progress 8 scores. The correlations in Table 3 show that in each case there was only a relatively weak relationship between the percentage of the ethnic group and Progress 8 mean scores. The correlation was negative between the percentage of White students and Progress 8 score (-0.23) whilst correlations were positive between the percentages of Asian, Chinese, or Black students and Progress 8 score (0.18, 0.17, and 0.14 respectively). Again, these results reflect the student-level results in Table 1, which show more progress on average made by Asian, Chinese or Black students, compared with White students.

Regression analysis of student-level Progress 8 scores

Table 4 presents the results of the modelling of school-level Progress 8 scores. Model 1 included no predictors, just an intercept, to assess the amount of variance in achievement between schools. From the random effects part of the table, we calculate that schools accounted for around
11.5 per cent of the variance in Progress 8 scores. This is a substantial proportion and suggests that the use of a multilevel model was justified. Model 2 then included all of the predictor variables, whether at the student or school level, with statistical significance (p<0.05) indicated in the table by bold type. A further model (results not shown in the table) included any statistically significant interactions between fixed effects. Some of the results from this final model are presented and discussed later in this article.

The size of each coefficient represents the change in Progress 8 associated with a particular category compared with the base category (for the categorical variables). As Progress 8 is essentially a mean grade measure, the coefficients represent the change in mean grade associated with a particular category. For continuous variables, the coefficients represent the change in mean grade associated with a unit increase in that variable.

**Main effects**

The results of Model 2 included a significant effect of prior attainment, which is surprising since the calculation of Progress 8 already accounts for this. The size of the effect (-0.23) suggests that an increase of one KS2 level for a student was associated with a fall in Progress 8 of between one fifth and one quarter of a grade. The existence of this effect was thought to be due to the presence of two different prior attainment effects, which cancel each other out overall (so that the mean Progress 8 score at each level of prior attainment is zero): firstly, a between-school effect, where schools with higher performing intakes tended to display higher Progress 8 scores (see Figure 2 and also the School KS2 mean coefficient), and secondly, a within-school effect where, within any given individual school, candidates with the highest prior attainment were likely to be those with the lowest Progress 8 scores.

This means that, for example, a high prior attaining student will tend to make more progress on average if they are in a school with other high prior attaining students than if they are amongst students with low prior attainment. Equally, a low prior attaining student will have higher Progress 8 scores on average if they are in a school with high-attaining students. If we compare high prior attaining and low prior attaining students in the same school, on average, the students with low prior attainment will have slightly higher Progress 8 scores.

The results also showed that females made more progress than males (0.25 of a grade on average). More disadvantaged students tended to make less progress: being an FSM student was associated with a significantly lower Progress 8 score (by 0.39 of a grade), and an increase in the level of implied deprivation was significantly associated with a fall in Progress 8 score (by 0.09, for each increase in IDACI of 0.1).

Furthermore, even after accounting for these effects at individual level, schools having a larger proportion of their students receiving free school meals were associated with lower Progress 8 scores. However, this effect was very small – an increase in FSM students of 20 per cent was associated with a fall in Progress 8 of 0.04. In terms of SEN, each SEN

---

6. As calculated by the intraclass correlation coefficient (ICC): ICC = school variance/(school variance + error variance) = 0.131/(0.131+1.008) = 0.115.
7. A full table of results, including interaction effects, can be found in the conference paper version of this article, available on the Cambridge Assessment website: http://www.cambridgeassessment.org.uk/cam-research/26-published-resources/conference-papers/
8. This within-school effect was confirmed by grouping candidates within each school into five groups based on their prior attainment. Each successively higher prior attaining group had a lower mean Progress 8 score (0.05, 0.04, 0.01, -0.01, and -0.08).

---

**Table 4: Student Progress 8 score - Regression coefficients**

Note: Standard errors in brackets; bold type indicates statistical significance at p<0.05.

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.012 (0.014)</td>
<td>0.027 (0.022)</td>
</tr>
<tr>
<td>KS2 Average Level</td>
<td>-0.226 (0.002)</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.249 (0.003)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-0.393 (0.003)</td>
<td></td>
</tr>
<tr>
<td>FSM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>-0.393 (0.013)</td>
<td></td>
</tr>
<tr>
<td>IDACI</td>
<td>-0.933 (0.013)</td>
<td></td>
</tr>
<tr>
<td>SEN status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>-0.480 (0.005)</td>
<td>-0.480 (0.005)</td>
</tr>
<tr>
<td>SEN support</td>
<td>-0.503 (0.011)</td>
<td>-0.503 (0.011)</td>
</tr>
<tr>
<td>Statement</td>
<td>-0.691 (0.022)</td>
<td>-0.691 (0.022)</td>
</tr>
<tr>
<td>EAL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>0.386 (0.006)</td>
<td>0.386 (0.006)</td>
</tr>
<tr>
<td>Yes</td>
<td>0.091 (0.034)</td>
<td>0.091 (0.034)</td>
</tr>
<tr>
<td>Unknown</td>
<td>-0.001 (0.001)</td>
<td>-0.001 (0.001)</td>
</tr>
<tr>
<td>School type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Academy (Comprehensive)</td>
<td>0.095 (0.040)</td>
<td>0.095 (0.040)</td>
</tr>
<tr>
<td>Academy (Selective)</td>
<td>-0.051 (0.036)</td>
<td>-0.051 (0.036)</td>
</tr>
<tr>
<td>Academy (Modern)</td>
<td>-0.039 (0.012)</td>
<td>-0.039 (0.012)</td>
</tr>
<tr>
<td>Comprehensive</td>
<td>0.066 (0.070)</td>
<td>0.066 (0.070)</td>
</tr>
<tr>
<td>Grammar</td>
<td>-0.013 (0.048)</td>
<td>-0.013 (0.048)</td>
</tr>
<tr>
<td>Secondary Modern</td>
<td>-0.113 (0.048)</td>
<td>-0.113 (0.048)</td>
</tr>
<tr>
<td>School gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>0.112 (0.028)</td>
<td>0.112 (0.028)</td>
</tr>
<tr>
<td>Boys’</td>
<td>0.022 (0.024)</td>
<td>0.022 (0.024)</td>
</tr>
<tr>
<td>Girls’</td>
<td>-0.001 (0.001)</td>
<td>-0.001 (0.001)</td>
</tr>
<tr>
<td>School KS2 mean</td>
<td>0.232 (0.043)</td>
<td>0.232 (0.043)</td>
</tr>
<tr>
<td>FSM percentage</td>
<td>-0.002 (0.001)</td>
<td>-0.002 (0.001)</td>
</tr>
<tr>
<td>IDACI mean</td>
<td>0.165 (0.144)</td>
<td>0.165 (0.144)</td>
</tr>
<tr>
<td>SEN percentage</td>
<td>0.002 (0.001)</td>
<td>0.002 (0.001)</td>
</tr>
<tr>
<td>EAL percentage</td>
<td>0.003 (0.001)</td>
<td>0.003 (0.001)</td>
</tr>
<tr>
<td>Asian percentage</td>
<td>-0.004 (0.001)</td>
<td>-0.004 (0.001)</td>
</tr>
<tr>
<td>Black percentage</td>
<td>0.000 (0.002)</td>
<td>0.000 (0.002)</td>
</tr>
<tr>
<td>White percentage</td>
<td>-0.001 (0.001)</td>
<td>-0.001 (0.001)</td>
</tr>
</tbody>
</table>

**Region**

<table>
<thead>
<tr>
<th>Region</th>
<th>Level 1 (Pupil level)</th>
<th>Level 2 (School level)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.008</td>
<td>0.895</td>
</tr>
<tr>
<td>York</td>
<td>0.131</td>
<td>0.085</td>
</tr>
</tbody>
</table>

---

© UCLES 2018
status was associated with less progression, with the biggest effect being for those with EHCP (-0.691).

Compared to White students, all other ethnic groups were associated with significantly more progress, even after accounting for other factors, such as implied deprivation. The largest effect was for Chinese students (0.43 of a grade), followed by students from an Other ethnic background (0.32), Black students (0.23), and Asian students (0.19). Having English as a second language was associated with better progress (by 0.39 of a grade), compared with students who spoke English as their first language. There were some significant school type effects, with being in a Selective academy associated with significantly higher Progress 8 scores (by 0.10 of a grade) and being in a non-academy Secondary Modern associated with significantly lower Progress 8 scores (by 0.11 of a grade), compared with being in a Comprehensive academy. There was also a statistically significant negative effect of attending a Comprehensive school, but this was very small. It is interesting that the effect of attending a Grammar school (either an academy or not) was much reduced compared to that found in the descriptive analysis (see Table 1). In fact, attending a non-academy Selective school was not significantly different from a Comprehensive academy. This is probably because the effect seen in the descriptive analysis was more due to attending a school with a high mean prior attainment than to attending a Selective school.

Students in single-sex boys’ schools made significantly more progress than those in mixed schools, even after accounting for the fact that these schools tended to have a higher attaining intake and were more likely to be selective than mixed gender schools. However, whilst the descriptive analysis showed that students in girls’ schools made the most progress on average, there was no significant difference between girls’ school students and mixed school students in the statistical model. This is likely to be because the effects of an individual’s gender and of school mean prior attainment were more important in determining Progress 8 scores.

There were several statistically significant school-level effects, but these were very small. These included the percentages of SEN, EAL and Asian students in a school. The coefficients for these effects ranged from -0.004 to 0.003. An effect of -0.004 meant that an increase of 20 per cent was associated with a reduction in the Progress 8 score of just 0.08.

Finally, there were significant effects of the geographical region of the school. Compared with London, being in any another region was associated with lower Progress 8 scores. However, the effects were smaller than those seen in the descriptive analysis (and were not significant for some regions). Even so, there is evidence that the London effect was present, even after accounting for other factors.

**Interaction effects**

There were a number of significant interaction effects in the final model (results not shown in Table 4). A few of the more interesting effects are presented in this section. For the complete set of results, please refer to the full paper on the Cambridge Assessment website. Using the values of the coefficients from the final model, it was possible to calculate predicted Progress 8 scores for different groups of students, and therefore demonstrate the effects of these interactions. These are shown in Figures 3 to 7.

**Gender x SEN status**

Figure 3 presents predicted Progress 8 scores for male and female students, with different levels of SEN status. This shows that female.
attending schools with different levels of mean prior attainment,

EAL x School KS2 mean

Figure 6 presents the predicted Progress 8 scores for students from different ethnic backgrounds, by school KS2 mean.

EAL x Ethnic group

Figure 5 presents the predicted Progress 8 scores for students from different ethnicities, by their EAL status. This shows that the effect of having EAL was different for each ethnicity. The effect was significantly larger for White students than any other ethnicity, and was smallest for Asian students.

Ethnic group x School KS2 mean

The effect of school prior attainment for students from different ethnic backgrounds is shown in Figure 6. This demonstrates that students from Chinese, Black, or Other backgrounds all had scores that were less sensitive to changes in school-level prior attainment, compared with White students. The biggest difference was for Chinese students, for whom school prior attainment had essentially no effect on their predicted Progress 8 score.

EAL x School KS2 mean

Figure 7 presents the predicted Progress 8 scores for students attending schools with different levels of mean prior attainment, by their EAL status. This shows that the effect of school-level prior attainment was much less for students with EAL, than for English speakers.

Discussion

Many of the findings presented in this article match those of previous research that investigated progress in secondary schools amongst different groups of students. We found that female students made on average a quarter of a grade (per qualification) more progress than males. This advantage in terms of progress in secondary school for female students was similar to that found in previous research (e.g., Sammons, 1995; Burgess, McConnell, Propper & Wilson, 2004). One possible explanation for this is the fact that, in the past, GCSEs had a substantial proportion of coursework or controlled assessment, which tend to favour girls. More recent research (Bramley, Vidal Rodeiro, & Vitello, 2015) has shown that girls also outperform boys in written examinations at GCSE, but by not as much as in coursework. It will be interesting to see whether the introduction of reformed GCSEs (most of which will no longer have coursework) leads to the gender gap closing somewhat.

We also found that students eligible for free school meals made less progress than other students, by around 0.4 of a grade on average. Students experiencing higher levels of implied deprivation also made less progress. Previous studies (e.g., Sammons, 1995; Strand, 2014; Sammons et al., 2014) found a similar negative relationship between progress and higher levels of disadvantage. One proposed reason was that more disadvantaged students have less support at home and lower educational aspirations.

The next finding was that all ethnic groups made more progress than White students, even after accounting for other factors, such as implied deprivation. The largest effect was for Chinese students (around 0.4 of a grade on average). Black and Asian students made about 0.2 of a grade more progress on average. This fits in with the findings from previous research (e.g., Sammons, 1995; Wilson, Burgess and Briggs, 2011; Strand, 2014), which suggests that the importance of high aspirations in many minority ethnic families and communities is the most likely explanation for this better progress.

We also found that students with EAL made about 0.4 of a grade more progress than students with English as their first language. Again, this matches the findings of previous research (e.g., Strand et al., 2015). However, part of this effect may be due to a change in the true EAL status over time. The NPD defines EAL students as those exposed to an additional language in their home or community, taking no account of their actual proficiency in English. We know that all of the students included in the analysis have been in the country for at least five years, because we have access to their KS2 test results. It may be that for many of these students, their knowledge of English will have improved significantly in that time and this may explain why their progress was so much greater.

Many of the school-level effects from the regression models were very small, but there were three variables (all closely linked) which were important. In the descriptive analysis we found that students in Selective schools or in single-sex schools made considerably more progress on average. However, in the regression model some of those differences disappeared, or became much smaller. In particular, the fact that students in schools with a higher attainment intake (as measured by KS2 mean) made more progress led to a big fall in the effect of attending a Selective school. Even so, some school type effects were still present after accounting for school mean KS2, with students in Selective academies performing better by about one tenth of a grade, and those in non-academy Secondary Moderns performing worse (by 0.11 of a grade) than students in Comprehensive academies. The girls’ school advantage disappeared completely in the statistical model, whilst the boys’ school advantage was still present.
This suggests that boys perform better in single-sex schools than in mixed schools, but girls do not, after accounting for other factors.

The region in which students attended school also had a significant effect on Progress 8 scores, with lower mean scores for all regions compared with London. However, the effects seen in the regression model were smaller than those in the descriptive analysis (and were not significant for some regions). For example, the North East had the second lowest mean Progress 8 score (-0.14) and was also highlighted in a recent DfE report (DfE, 2017b) as having the highest percentage of schools below the ‘floor standard’. However, this effect was not significant once other factors were taken into account. Despite these caveats, there was still evidence that the London effect was present, even after accounting for other factors.

Some interesting interaction effects were also present. The advantage for girls was much smaller (or disappeared completely) if they had any SEN. The implied deprivation effect was greatest for White students, compared with other ethnic backgrounds. Figure 4 showed that White students from more deprived backgrounds made the least progress on average. This confirms the findings of previous research (e.g., Strand, 2014), and suggests that the importance of high aspirations in non-White ethnic backgrounds extends to those from poorer families.

Perhaps of most importance was the finding that, despite the measure taking account of prior attainment at the student level, Progress 8 systematically penalised schools with a lower performing intake. This pattern was present for both Selective and non-Selective schools, which suggests that the difference seen between school types was more due to the prior attainment of the students than to the type of school per se. This confirms previous findings (e.g., Yang & Woodhouse, 2001) that students make more progress in schools with a higher attaining intake. It is not entirely clear why this would be the case, but Thomson (2015) has suggested three possible explanations: students in higher attaining schools receive more help at home; competition amongst students may drive up attainment; and higher attaining schools may be more effective because they can recruit better teachers. It may also be the case that low prior attaining students receive a boost in high-attaining schools due to the positive influence of higher attaining peers.

Overall, this analysis has shown that Progress 8 scores are strongly related to a plethora of school and student factors, most of which are beyond the control of teachers. Therefore, it is questionable that Progress 8 should be used as a measure for comparing schools, or for bringing them to the attention of the Office for Standards in Education, Children’s Services and Skills (Ofsted) if they are below the floor standard. It could be argued that a fairer way of judging schools would be to take account of some of these factors when calculating school performance measures. Whilst debate over different ways of measuring school effectiveness is likely to continue (e.g., Allen, 2015; or Allen, Burgess & Mayo, 2018), an awareness of the potential limitations of any such measures is crucial to their interpretation.

References


Postgraduate Certificate in Educational Assessment and Examinations

Take your understanding of assessment to the next level

The Postgraduate Certificate in Educational Assessment and Examinations is a practice-based qualification designed to directly impact your work as you learn to apply various research methodologies to your professional context.

It is taught through a mix of online learning and Saturday Day Schools in Cambridge and led by experts from Cambridge Assessment and the University of Cambridge Faculty of Education.

You will be awarded a Postgraduate Certificate in Educational Studies (PCES) worth 60 credits at Master’s level (Level 7) at the end of the course.

“It was a very valuable learning experience that I will be returning to repeatedly over the next 12 months, re-reading, re-thinking and adapting practice.”

Find out more: www.cambridgeassessment.org.uk/events/pgca/
Conference presentations

Educational Collaborative for International Schools (ECIS) Leadership Conference

American Educational Research Association (AERA)
The AERA conference took place in New York, USA, in April 2018 with the theme The Dreams, Possibilities, and Necessity of Public Education. Carmen Vidal Rodeiro, Research Division, presented Identifying ‘Fit-for-Purpose’ Data for Predictive Validity Studies to Inform Postsecondary Admissions Decisions, a paper co-authored with Stuart Shaw.

The European Conference on Education (ECE)

International Conference on Education and New Learning Technologies (EDULEARN)
The 10th annual EDULEARN conference was held in Palma de Mallorca, Spain, in July 2018. The event is a platform for international strategic networking and for presenting innovations and projects about education and technology. Stuart Shaw presented the following co-authored papers:

A seminar with Dr Mary Richardson, UCL Institute of Education
This seminar will consider how assessment is discussed and perceived within the "post-truth" perspective applied to education.

Register today:
www.canetwork.org.uk/seminars
29 January 2019 | 16.00 – 18.30 | Cambridge | Free
Stuart Shaw and Marta Czajkowska, Cambridge Assessment International Education: The issues in the design, construction and administration of student questionnaires for bilingual learners.

Stuart Shaw and Carmen Vidal Rodeiro: The value of predictive validity studies and the need for ‘fit-for-purpose’ data to inform postsecondary admissions policies and decision-making in the United States.

5th International Conference on Employer Engagement in Education and Training

Held in London, UK, in July 2018, the Education and Employers and the Edge Foundation brought together leading researchers, practitioners, and policy makers from around the world to present recent research and discuss employer engagement in education, policy development and delivery. The theme of the event was Social Mobility, industrial and skills strategies, post-Brexit: What role can employer engagement and technical education play? Several researchers from the Research Division of Cambridge Assessment attended the conference and the following papers were presented:

Sylvia Vitello and Simon Child: Vocational qualifications for 14–16 year olds: Exploration of knowledge, skills and teacher perceptions.

Joanna Williamson and Matthew Carroll: Examining the use of technical qualifications within Key Stage 5 programmes of study.

Further information on all conference papers can be found on our website: http://www.cambridgeassessment.org.uk/our-research/all-published-resources/conference-papers/

International Summit on Textbooks

This event was held in London, UK, June 2018 and was hosted by The Rt. Hon. Nick Gibb, UK Minister of State for School Standards, in collaboration with Cambridge Assessment and the Royal Society. It enabled a high-quality debate on internationally experienced views on textbooks, from jurisdictions that either routinely use textbooks or are interested in introducing them to improve education standards.

Tim Oates, CBE, Cambridge Assessment’s Group Director of Assessment Research and Development, spoke about research on the strategic role of high-quality materials. Issues explored at the summit included changes in the development and use of textbooks, how teachers can use textbooks to reduce workload, and the incentives and barriers to promoting high-quality textbooks. Further insight from the day can be found on our website: http://www.cambridgeassessment.org.uk/news/a-textbook-a-day/

Publications

The following articles have been published since Research Matters, Issue 25:


Further information on all journal papers and book chapters can be found on our website: http://www.cambridgeassessment.org.uk/our-research/all-published-resources/journal-papers-and-book-chapters/

Data Bytes

Data Bytes is a series of data graphics from Cambridge Assessment’s Research Division, designed to highlight the latest research findings and trends in educational assessment to a wide audience. Topics are often chosen to coincide with contemporary news or recent Cambridge Assessment research outputs. All Data Bytes can be found on our website: http://www.cambridgeassessment.org.uk/our-research/data-bytes

The following Data Bytes have been published since Research Matters, Issue 25 – Interactive graphics are indicated with (I):

- May 2018: Preparing students for university (I)
- June 2018: Cambridge Assessment exam papers shipping (I)
- August 2018: Top GCSE news stories (I)
A101: Introducing the Principles of Assessment

An interactive online course designed to provide you with an accessible but thorough grounding in the principles of assessment.

- No previous knowledge or experience needed
- Average 3 hrs per week over nine weeks
- Certification option
- Moderated by assessment experts with weekly video summaries
- Monitored option: £275 per person
  Peer supported option: £225 per person

A101: Introducing the Principles of Assessment is an online course created by the Cambridge Assessment Network for anyone with an interest in educational assessment and its role in society today.

The course covers validity, reliability, fairness, standards, comparability, practicality and manageability of assessment.

To find out more and book a place, visit: www.canetwork.org.uk/a101
Contents / Issue 26 / Autumn 2018

2 To "Click" or to "Choose"? Investigating the language used in on-screen assessment: Rushda Khan and Stuart Shaw
9 Articulation work. How do senior examiners construct feedback to encourage both examiner alignment and examiner development? Martin Johnson
15 Characteristics, uses and rationales of mark-based and grade-based assessment: Joanna Williamson
22 Is comparative judgement just a quick form of multiple marking? Tom Benton and Tom Gallacher
28 How have students and schools performed on the Progress 8 performance measure? Tim Gill
38 Research News - Karen Barden