Cambridge Assessment

Research *Matters*

Issue 24 / *Autumn 2017*

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

This publication should be cited using the following example for article 1:
Darlington, E. and Bowyer, J. (2017). Undergraduate Mathematics students' views of their pre-university mathematical preparation. *Research Matters: A Cambridge Assessment publication*, *24*, 2–11.

Credits

Editorial and production management:
Tom Bramley, David Beauchamp and Karen Barden, Research Division, Cambridge Assessment
Additional proofreading: Jo Ireland and Victoria Coleman, Research Division, Cambridge Assessment
Cover image: iStock.com/Ivanastar
Design: George Hammond
Print management: Canon Business Services

1 **Foreword** : Tim Oates, CBE

1 **Editorial** : Tom Bramley

2 **Undergraduate Mathematics students' views of their pre-university mathematical preparation** : Ellie Darlington and Jessica Bowyer

11 **Question selection and volatility in schools' Mathematics GCSE results** : Cara Crawford

17 **Utilising technology in the assessment of collaboration: A critique of PISA's collaborative problem-solving tasks** : Stuart Shaw and Simon Child

23 **Partial absences in GCSE and AS/A level examinations** : Carmen Vidal Rodeiro

30 **On the reliability of applying educational taxonomies** : Victoria Coleman

37 **How much do I need to write to get top marks?** : Tom Benton

42 **Research News :** David Beauchamp, Karen Barden, Gill Elliott, and Gillian Cooke

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

The full issue of *Research Matters 24* and all previous issues are available from our website: www.cambridgeassessment.org.uk/research-matters

Research *Matters* / 24

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

This issue of *Research Matters* makes the national scene in England appear calmer than really it is. Calm it is not. I write this foreword now, in June 2017, knowing it is likely that this issue will be read for many years to come, by readers who are not aware of the gravity of the changes to public examinations which are impacting this year. Cambridge Assessment played a key role in the development of the revised National Curriculum, which was designed from 2010–13 and first implemented in schools in September 2014. Alongside these changes, the wide set of reforms included revised GCSEs and A levels, these latter – following Cambridge Assessment's recommendations – being linked more closely to the entry requirements of Higher Education; the former elevated in demand, with a new grade structure – 9-1 rather than A*-G. Concerned that GCSEs had fallen behind the standards emerging globally in high-performing jurisdictions, Secretary of State Michael Gove decided on a demanding policy option. Knowing that it is GCSE and not the National Curriculum which dominates school learning from age 14–16, change in the content of the examinations was seen as a critical priority. The simplest policy option – to leave the qualifications unchanged but to move from C to B as the lowest recognised 'target grade' – was eschewed in favour of root and branch reform of content, and a new grade structure (lowest 1, highest 9) was introduced to give greater differentiation at the top of the scale, with '5' as the new lowest 'target grade' expectation – the lower '4' corresponding to the old 'C' grade. Cambridge Assessment has of course provided a suite of exams alongside the other exam boards, but its policy work resulted in a revised national model of assessment of practical work in Science GCSE and A level, and to the national model for awarding the top '9' grade in GCSE. First awards are occurring this summer (2017) and all stakeholders – parents, candidates, schools, government, exam boards and the national regulator understandably are watchful and anxious. The stakes are high for them all, but for different reasons in each case. Using 'comparable outcomes' as an approach to smooth the transition into the new qualifications provides some assurance in general, but until all results are known, current speculation about the way in which individual pupils and schools in different localities have adapted to the new demands will be conjecture rather than fact.

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

Editorial

Mathematics is the most popular A level subject, and it has been reported that it can add (on average) 11 per cent to your salary by age 34[1]. But how useful is it as preparation for an undergraduate degree in Mathematics? The first article in this issue by Darlington and Bowyer reports on a survey of undergraduate mathematicians on how well A level Maths and Further Maths had prepared them for their university studies. They consider the impact of A level reform in terms of the shift from modular to linear exams and the introduction of compulsory Mechanics and Statistics content. The second article by Crawford continues a strand of work in the Research Division aiming to understand the causes of volatility in schools' exam results. The third article by Shaw and Child considers the problems arising from the attempt to assess collaboration, and evaluates the extent to which technology can provide a solution. The fourth article by Vidal Rodeiro explores a range of methods from simple to complex for estimating 'missing marks' on components of GCSE and A level exams that arise (for example, when illness prevents a student from sitting the exam).

Specifying the knowledge and skills being assessed is an important aspect of test validity, and several classification schemes or taxonomies have been developed for this purpose. However, it is difficult to unambiguously classify test questions, and experts inevitably differ, so it is important to assess the reliability of classification when such taxonomies are applied. Coleman reviews some of the literature and finds that the majority of studies have reported moderate to good reliability. Finally, Benton attempts to shed some light on the age-old question of 'how much should I write?', using image processing techniques to identify the distinct words in scanned images of candidates' answers to an essay question. Reassuringly he finds that although there is a relationship between number of words written and mark obtained, quantity doesn't 'trump' quality!

1. Adkins, M., & Noyes, A. (2016). Reassessing the economic value of advanced level mathematics. *British Educational Research Journal*, *42*(1), 93–116.

Tom Bramley *Director, Research Division*

Undergraduate Mathematics students' views of their pre-university mathematical preparation

Ellie Darlington Research Division and **Jessica Bowyer** Exeter University (The study was completed when the second author was based in the Research Division)

Introduction

This research took place during an extensive A level reform programme conducted by the UK government's Department for Education, in order to inform the redevelopment of the content and assessment used as part of optional Further Mathematics units. Whilst much research has been conducted on the transition between secondary and tertiary Mathematics in the UK and internationally, research has as yet not been conducted regarding students' perceptions of the usefulness of A level Mathematics and Further Mathematics. This study sought to answer the following research questions:

1. Which optional units in A level Mathematics and/or Further Mathematics did students find useful as preparation for their degree?
2. Did students believe that A level Mathematics and Further Mathematics were useful preparation for their degree?
3. Are there any areas in which A level Mathematics and/or Further Mathematics could be improved to suit the needs of future prospective Mathematics undergraduates?

Whilst this article reports on the responses of Mathematics undergraduates, the data was collected as part of a larger overarching project which sought the views of over 4,000 undergraduate Science and Social Science students regarding their perceptions of A level Mathematics as preparation for the mathematical demands of their degree (Darlington & Bowyer, 2016).

Mathematics degree courses

In recent years, the number of Mathematics undergraduates has increased substantially, both in absolute numbers and in terms of the proportion of Mathematics students of the whole full-time undergraduate population (see Figure 1). Numbers have increased from 13,188 (1.3% of all full-time undergraduates) in 1996/97 to 27,810 (2.1%) in 2014/15. The number of Mathematics graduates is important for economic development, and thus the need for a large number of mathematically-competent graduates continues to increase (Gago, 2004; Petocz & Reid, 2005; Wolf, 2002).

A levels

Advanced, or 'A' level qualifications in England and Wales are post-compulsory qualifications taken at the end of secondary schooling at age 18. A wide variety of subjects are available for students to choose from, with most studying three or four subjects over a two-year period. Students are then awarded separate grades for each subject. Whilst there are no compulsory A level subjects, A level Mathematics was the most popular subject in 2016, comprising 11.0% of all A levels taken (Joint Council for Qualifications [JCQ], 2016). An A level is also available

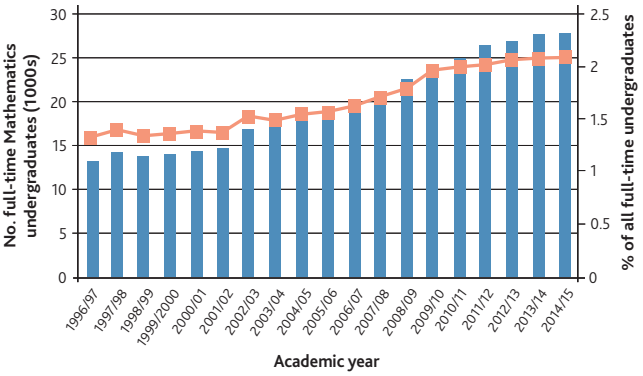

Figure 1: Full-time students of Mathematics degrees in the United Kingdom, 1996-2015 *Data from the Higher Education Statistics Agency (1998-2016).*

in Further Mathematics, which is one of the fastest-growing subjects in terms of uptake.

At the time of writing, A levels are examined at the end of a two-year course, with separate assessment for each unit which makes up each A level. Students are able to be examined at the end of the first year of the course, earning an Advanced Subsidiary 'AS' level. Students are generally required to achieve certain grades in A levels to be accepted onto a university degree course.

A level Mathematics and Further Mathematics: pre-September 2017

A level Mathematics comprises six equally weighted units (modules), which are individually assessed in 90 minute examinations. Four of these units are compulsory Pure Mathematics units: 'Core Pure Mathematics'. Two optional units may then be chosen from three Applied Mathematics strands, each of which contains up to five sequential units: Mechanics (M1-M5), Statistics (S1-S4), and Decision Mathematics (D1-D2). The two optional units may be chosen from the same strand (e.g., Statistics 1 + Statistics 2) or from a mixture of two (i.e., M1 + S1, M1 + D1, S1 +D1). Consequently, there are six routes through A level Mathematics. AS level Mathematics comprises Core Pure Mathematics 1 (C1) and Core Pure Mathematics 2 (C2) and one of S1, D1 and M1.

A level Further Mathematics comprises three compulsory Further Pure (FP) Mathematics units, and three optional units. The optional units may be selected from any combination of FP and Applied options from the three strands which are available as part of A level Mathematics. Students may not take units in AS or A level Further Mathematics which they have already taken as part of A level Mathematics.

Whilst Further Mathematics has been found to be the most (relatively) demanding A level subject (Hillman, 2014), high numbers of students achieve high grades in the subject – in 2016, 56.2% of A level Further Mathematics students achieved at least an A grade (JCQ, 2016).

A level Mathematics and Further Mathematics: post-September 2017

The A level system is currently undergoing a general reform programme with specific changes also planned to take place within certain subjects. From September 2017, there will be no optionality in A level Mathematics. All students will study Pure Mathematics, Mechanics and Statistics content. This change will mean that university admissions tutors will know that all students who have taken A level Mathematics will definitely have studied certain content. At present, for example, students embarking upon Engineering degrees (where prior study of Mechanics is beneficial) may have studied up to five units from the Mechanics strand, depending on the units they studied and whether or not they studied Further Mathematics.

Decision Mathematics will no longer be available for study as part of A level Mathematics, though it may be available through the study of Further Mathematics. Half of Further Mathematics' content will be compulsory Pure Mathematics material, with the remaining half optional. Optional content will be decided by the awarding bodies of the examinations, and is likely to involve innovative new units which might not necessarily follow the currently Decision-Mechanics-Statistics structure.

A level Mathematics as preparation for undergraduate mathematics

Apart from A level Mathematics and Further Mathematics, the most popular A level subjects taken by new Mathematics undergraduates are predominantly in the Sciences or Computing. These A levels also have mathematical components (see Table 1).

Table 1: Top 10 A level subjects taken by 2011's new undergraduate mathematicians (Vidal Rodeiro & Sutch, 2013)

Rank	*Subject*	*% students*
1	Mathematics	68.2
2	Physics	32.4
3	Further Mathematics	30.9
4	Chemistry	19.1
5	Information Communication Technology	12.4
6	General Studies	11.9
7	Computing	11.5
8	Biology	11.4
9	Economics	10.0
10	Business Studies	9.3

Note: This is the most recently available data.

As well as a rising number of entrants to undergraduate Mathematics, an increasing number of students are taking post-compulsory Mathematics qualifications. Mathematics is the most popular subject at A level, with 92,163 candidates in 2016 (JCQ, 2016). Additionally, Further Mathematics is the fastest growing A level, with the number of students taking this qualification more than doubling from 5,720 candidates in 2004 to 15,257 in 2016 (JCQ, 2016). The increasing numbers of students choosing to study post-compulsory Mathematics has been attributed to the changing economic climate. That is, students realise that Mathematics has a high exchange value in the workplace and in Higher Education, and therefore they study it in order to increase their future job prospects.

Mathematics requirements for undergraduate courses in the UK

Students are required to achieve high grades in A level Mathematics in order to study the subject at most universities, and increasing numbers of universities are making Further Mathematics a compulsory entry requirement. In fact, the increases in the proportion of new undergraduates who have taken A level Further Mathematics mean that some universities are now changing the structure and content of the first year of their degrees to accommodate the changing mathematical backgrounds of their students (Searle, 2014).

Students in the Mathematical Sciences are more likely to achieve top grades at A level than students in other subject areas, with 8.1% of them achieving three or more A* grades at A level (Vidal Rodeiro & Zanini, 2015), a figure which is much higher than in other degree subjects. Similarly, a greater proportion of Mathematical Sciences students graduate with a First-class degree result (32%) than other subjects (Vidal Rodeiro & Zanini, 2015).

The Mathematics problem

The preparedness of British undergraduate mathematicians for the demands of university study has been of concern since the 1990s. The term 'Mathematics problem' is used widely to describe anxieties regarding the relatively small number of students choosing to study the subject at tertiary level, not just in the UK but on an international scale. This has sometimes been attributed to an increased number of students having negative experiences of Mathematics at school (Smith, 2004). Furthermore, once students advance to undergraduate study, many fail to succeed in this new environment – the Mathematical Sciences had the highest drop-out rate (24.0%) of all disciplines in 2014/15 (Higher Education Statistics Agency [HESA], 2016a).

Savage (2003) reported that this phenomenon occurred despite many Mathematics students achieving good grades at A level, something which is essential for students to be accepted onto Mathematics degree courses. According to Savage (2003), incoming students were lacking in three areas:

1. They were unable to fluently and consistently perform algebraic manipulations and simplifications
2. Their analytical powers were weak in instances where they were required to solve multi-step problems
3. They were ignorant of the nature of Mathematics and, more specifically, undergraduate Mathematics.

Concerns have been raised over the past few decades that new students arrive at university with insufficient Mathematics knowledge (ACME, 2011; Williams, 2011). The skills taught at school are often considered by universities to be an insufficient basis for further study in Mathematics, and the gap between secondary and tertiary is widely researched and debated, with Tall (1991) describing it as a shift "from *describing* to *defining*, from *convincing* to *proving* in a logical manner based on those definitions" (p.20). The mathematical competency of incoming undergraduates has been found by Smith (2004) to be decreasing over time, with scores on a diagnostic test for new students decreasing with each cohort.

This frequently manifests itself in students' difficulties with mathematical proof. Selden (2012) called the new emphasis on proof at the undergraduate level a 'major hurdle' for newcomers, with much of it centred on mathematical analysis.

Criticisms of the current secondary curriculum and assessment regarding the preparedness of new undergraduate mathematicians have resulted in the evolution of the university Mathematics curriculum. The apparent discrepancy between what students actually know post-A level and what their lecturers expect them to know when they begin university study "will, at the very least, impair the quality of their education and, at worst, may prove too difficult for them to bridge" (Lawson, 1997, p.151).

As the content (and purpose) of A level Mathematics has continued to change throughout the decades, universities have made a number of concessions to change. Consequently, diagnostic testing is now used in many Mathematics departments across the UK (Edwards, 1996; LTSN MathsTEAM, 2003; Williams, Hernandez-Martinez, & Harris, 2010), with many universities conceding that "the idea that the final year of school should fit the students for the first year of mathematics is no longer automatic" (Baumslag, 2000, p.6).

The impact of mathematical backgrounds on degree performance

In a study of the secondary-tertiary Mathematics interface, Kajander and Lovric (2005) found that students' school experiences often shape study approaches at undergraduate level. These stemmed from their beliefs about Mathematics which were that Mathematics is a rule-based subject requiring the learner to memorise facts and algorithms (Anderson, Austin, Barnard, & Jagger, 1998; Crawford, Gordon, Nicholas, & Prosser, 1994, 1998a, 1998b).

The difference between secondary and undergraduate Mathematics in the UK has been outlined by Darlington (2014), who used the Mathematical Assessment Task Hierarchy (Smith et al., 1996) to compare the types of skills required to answer questions in A level Mathematics and Further Mathematics, admissions tests, and undergraduate Mathematics examinations. This analysis revealed A level to be dominated by the routine use of procedures, but undergraduate examinations to emphasise proofs and interpretations.

Indeed, Gueudet (2008) argues that, at school, "students just have to produce results. At university, they seem to have an increasing responsibility towards the knowledge taught" (p.240). This takes the form of applying what they have been taught in a creative fashion which should ultimately allow them to construct proofs of mathematical statements and conjectures; however, many have a "(false) belief that, given sufficient time and study, there will be an algorithm that will solve any given problem" (Ervynck, 1991, p.52). Students' ability to apply what they have learnt at school in terms of their mathematical understanding, learning approaches and conceptions of Mathematics to the undergraduate setting is essential in their success with the subject at tertiary level (Wood, 2001). Consequently, many students report experiencing a 'bump' in their educational path (Perrenet & Taconis, 2009).

Method

An online questionnaire was developed in order to gain an insight into Mathematics undergraduates' perceptions of their mathematical preparedness. Only students who had taken A level Further Mathematics were eligible to take part, as students were asked specifically to reflect on how well A level Mathematics and Further Mathematics had prepared them. They must also have completed at least one year of degree study in order that they could reflect on their experiences so far.

Students from different universities are subjected to different admissions requirements, and study different content, receive different types of teaching, and are subjected to different examination and assessment systems. The questionnaire was publicised to all UK universities offering single honours Mathematics degrees in the hope that participation could be gathered from a wide cross-section of the student population.

The questionnaire was developed by the authors in conjunction with A level Mathematics qualification specialists, and piloted with three recent graduates of STEMM and Social Science degrees who had taken A level Mathematics and Further Mathematics to ascertain whether the questions were appropriate, effective and clear. Minor changes were made in response to the piloters' feedback. The questionnaire sought to survey participants regarding:

- Mathematical background
- Current study and performance
- Perceptions of mathematical preparedness.

Results

The results of statistical testing in this article all refer to chi-squared tests, or Fisher's exact test, where a chi-squared test could not be performed.

Sample

After data cleaning, 928 students participated in the study.

Gender: The sample consisted of 35.6% female and 63.4% male participants. This is representative of the ratio of males to females studying Mathematics at university; in the 2014/15 academic year, 62% of undergraduates in the Mathematical Sciences at British universities were male (HESA, 2016b).

Study institution: Participants came from 42 different universities, with a median of 65 per university. Most participants studied at universities in England (91.9%), with 2.3% in Scotland, 2.1% in Northern Ireland and 3.7% in Wales.

Degree programmes: Students participating in the questionnaire studied one of 46 different specific degrees, including joint honours degrees (see Table 2). Most participants were in their second year of study (50.8%), with 34.5% in their third and 14.7% in their fourth.

Table 2: Degree courses studied

	Degree	*No. participants*	*% participants*
Single honours[a]		690	75.1
Joint honours	Physics	43	4.7
	Economics	40	4.4
	Operational Research	32	3.5
	Computer Science	30	3.3
	Statistics	29	3.2
	Finance	16	1.7
	Philosophy	11	1.2
	Modern Foreign languages	8	0.9
	Business Studies	7	0.8
	Other	13	1.4
Total		**919**	**100**

[a] including degrees such as 'Mathematical Sciences' and 'Mathematical Studies'.

Table 3: Participants A level Mathematics and AS/A level Further Mathematics grades

Grade	*% Students*						
	A level Mathematics			*AS level Further Mathematics*		*A level Further Mathematics*	
	Participants (n=927)	*All Mathematics & Computer Science undergraduates (2011)*	*All candidates (2016)*	*Participants (n=112)*	*All candidates (2016)*	*Participants (n=788)*	*All candidates (2016)*
A*	72.5	34.7	17.5	N/A	N/A	53.4	28.7
A	22.3	29.1	24.3	60.7	53.8	27.4	27.5
B	4.3	17.3	22.3	21.4	16.7	12.4	20.6
C	0.6	10.3	16.1	8.0	11.6	4.9	11.3
D	0.2	5.8	10.8	4.5	7.5	1.6	6.5
E	0.0	2.7	6.1	3.6	5.0	0.1	3.5
Fail	0.0	0.1	2.9	1.8	5.4	0.0	1.9

Additional data from the JCQ (2016) and Vidal Rodeiro (2012).

Academic performance: The majority (87.2%) of participants had taken the full Mathematics A level, and 12.8% the AS level. Only 41.7% of participants were required to have taken AS or A level Further Mathematics to be accepted onto their course.

In both Mathematics and Further Mathematics, most participants achieved an A*. Table 3 shows that the average participant was therefore a higher attainer than the average A level Mathematics or Further Mathematics student in 2016. This is particularly the case for A level Mathematics.

Students who were required to have taken Further Mathematics to be accepted on their course were more likely ($p<.001$) to have achieved an A* in A level Mathematics (86.1%) and Further Mathematics (61.8%) than those who were not (53.7% and 24.9%, respectively).

Significantly more males were awarded higher grades in A level Mathematics than women were ($p<.001$), with 61.8% of women achieving an A* compared to 78.7% of men. However, the proportions of each gender achieving grade B or lower were very similar.

Most participants were awarded their final Mathematics qualification in 2013 (42.1%) or 2012 (34.3%), with some finishing their A levels as far back as 2006.

Of the 928 participants, 916 (98.7%) recalled taking at least one Mechanics unit, 902 (97.2%) a Statistics unit, and 676 (72.8%) a Decision Mathematics unit (see Figure 2). Most of those who took Decision Mathematics only took D1. However, most of those who studied Mechanics or Statistics took two units from those strands. Most participants studied Further Pure Mathematics up to FP3, reflecting that the majority had taken the full A level in Further Mathematics.

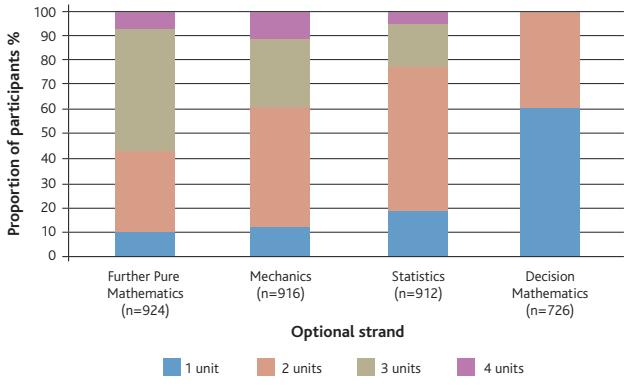

Figure 2: Number of optional units studied

Men studied significantly more Mechanics units than women ($p<.001$), though the majority of both male and female participants reported that they had studied two Mechanics units. Nearly 33% of men and 20.8% of women took three Mechanics units, with 11.1% of men taking four units compared to only 2.1% of women. Furthermore, 19.9% of women only studied M1, compared to only 8.2% of men.

Experiences of non-compulsory A level units

Participants were asked to comment on the relative utility of the non-compulsory units that they studied as part of Mathematics and Further Mathematics as preparation for university Mathematics (see Figure 3).

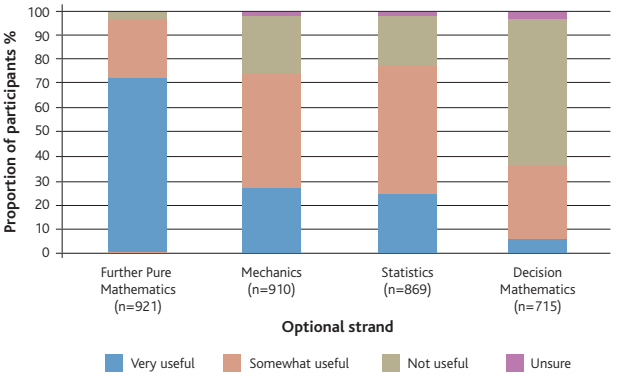

Figure 3: Students' views of the utility of optional units

Further Pure Mathematics, two units of which is compulsory in AS level Further Mathematics and three units for A level Further Mathematics, were described most positively. Overall, of the four strands participants were questioned about, these units were described most commonly as very useful preparation for their degree (73.0% participants). Only 22 of the 921 (2.4%) who answered this question described it as not useful.

Similar proportions of participants found Mechanics and Statistics units to be very or somewhat useful as preparation for their

undergraduate courses (74.5% and 77.4%, respectively). However, the proportion of participants who reported that Decision Mathematics units were useful preparation was much smaller, with 60.2% reporting that this strand was not useful. Conversely, only 23.5% of participants who took Mechanics and 28.4% of those who took Statistics described them as not useful.

Significantly more women than men ($p<.001$) perceived Statistics to be useful for their degree, with 39.8% of the former describing it as very useful but only 22.0% of the latter agreeing.

Motivations for studying Further Mathematics

Students were asked to indicate how influential certain factors were in their decisions to study Further Mathematics (see Figure 4) using statements which had been used in an earlier study regarding the uptake of A level Mathematics (Qualifications and Curriculum Authority, 2007).

The vast majority of participants (89.5%) reported that they were influenced in their decision to take Further Mathematics to some extent by their perception that they had coped well with GCSE Mathematics.

Of the 15 options given to participants, the three factors most influencing their decision to study Further Mathematics were:

1. **Enjoyment of school Mathematics:** 85.0% of participants reported that this influenced them a lot in their decision-making.
2. **Being better at Mathematics than at other subjects:** 95.5% of participants reported that this influenced them either a lot or a little in their decision to study Further Mathematics.
3. **Thinking of studying for a Mathematics or Mathematics-related degree:** Only 5.0% of participants reported that this had no bearing on their decision to study Further Mathematics.

Factors which had little impact on participants' decisions to choose Further Mathematics included encouragement by parents, their school Mathematics departments' results and the subject choices of their peers.

Experiences of Further Mathematics

Participants were asked to indicate their relative agreement with 10 statements regarding their experiences of studying Further Mathematics. The data indicates that participants were generally positive (see Table 4).

Responses indicate that participants generally enjoyed Further Mathematics and were glad that they had taken it. Encouragingly, considering the sample and their A level performance (see Table 3), more than 60% of participants indicated agreement with the statement 'I found Further Maths challenging', although only 36.2% reported that Further Mathematics was their most difficult A level.

However, there were significant gender differences in these aspects, with women much more likely to have found Further Mathematics challenging than men ($p<.001$) and more likely to agree that it was their most difficult A level ($p=.006$). The requirement of Further Mathematics for university entry also affected responses, with 27.2% of those who were not required to have taken it strongly agreeing that Further Mathematics was their most difficult A level, compared to only 15.5% of those who were required to have taken it.

Additionally, whether or not the participant was required to have taken Further Mathematics to be accepted onto their current degree course impacted responses. Only 54.1% of such students agreed or strongly agreed that Further Mathematics was challenging, compared to 84.3% of participants who were not required to have taken Further Mathematics.

Most participants (79.9%) reported that there was some overlap between what they had studied at A level and what they were taught in the first year of their degree. Perhaps indicating that universities tailor their courses well for the entry requirements they make of their students, only 30.3% of participants who were required to have taken Further Mathematics strongly agreed that they were taught Further Mathematics material in their first year of university, compared to 48.7% of those who were not required to have taken Further Mathematics.

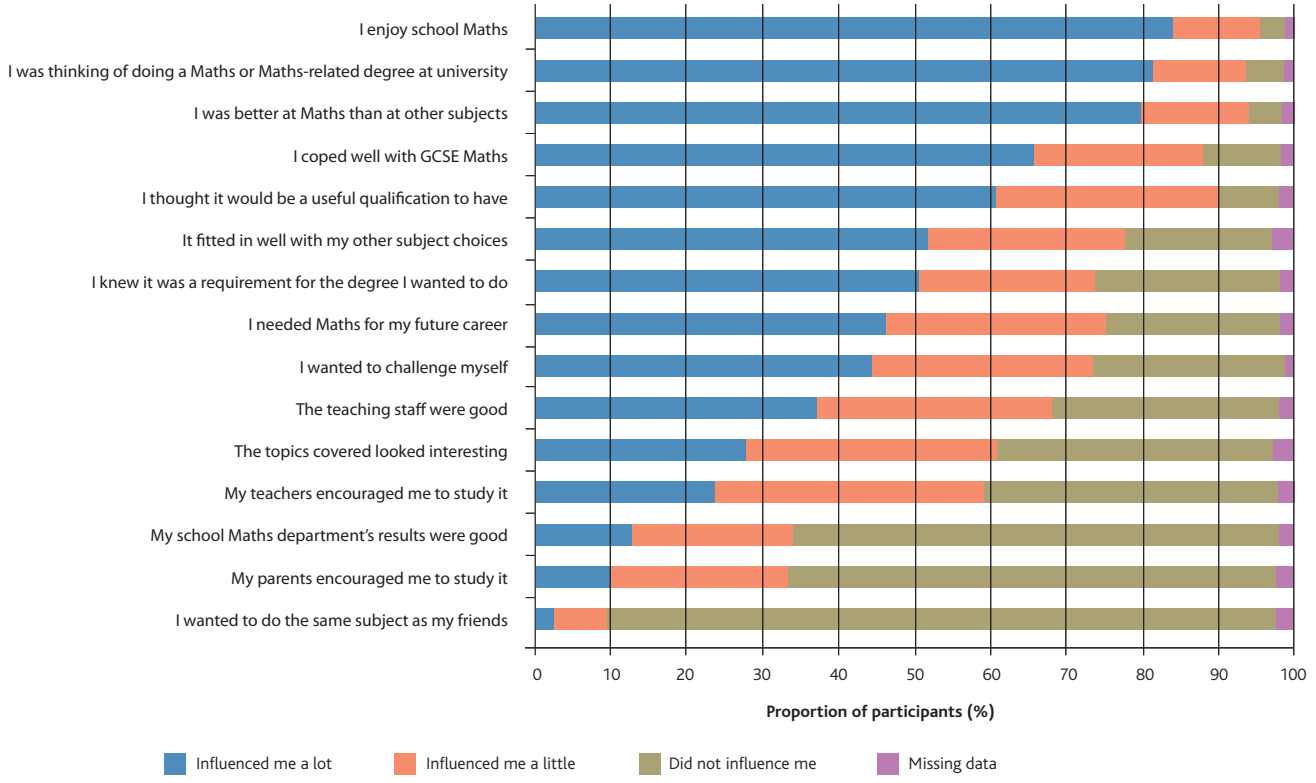

Figure 4: Students' motivations for studying Further Mathematics (*N*=928)

Table 4: Participants' experiences of studying Further Mathematics

	Strongly agree	*Agree*	*Neither agree nor disagree*	*Disagree*	*Strongly disagree*	*Unsure*
	No. participants (%)					
Further Maths was my most difficult A level	188 (20.4%)	143 (15.5%)	128 (13.9%)	299 (32.5%)	156 (16.9%)	7 (0.8%)
I'm glad I did Further Maths	673 (73.2%)	222 (24.1%)	19 (2.1%)	4 (0.4%)	2 (0.2%)	0 (0.0%)
I enjoyed Further Maths	492 (53.3%)	359 (38.9%)	50 (5.4%)	19 (2.1%)	3 (0.3%)	0 (0.0%)
In my first year at university, we were taught material that I had learned in Further Maths	349 (37.7%)	386 (41.7%)	76 (8.2%)	77 (8.3%)	32 (3.5%)	5 (0.5%)
Most people on my university course studied Further Maths	456 (49.5%)	193 (21.0%)	100 (10.9%)	96 (10.4%)	23 (2.5%)	53 (5.8%)
I found Further Maths challenging	211 (22.8%)	407 (44.0%)	143 (15.5%)	120 (13.0%)	43 (4.6%)	1 (0.1%)
Studying Maths and Further Maths was sufficient preparation for my degree	266 (28.9%)	351 (38.2%)	94 (10.2%)	137 (14.9%)	69 (7.5%)	2 (0.2%)
In my first year at university, we were taught material that I had learned in Further Maths	349 (37.7%)	386 (41.7%)	76 (8.2%)	77 (8.3%)	32 (3.5%)	5 (0.5%)

A levels as preparation for Mathematics degrees

Participants were asked how well they thought that A level Mathematics and Further Mathematics prepared them for their degrees. In the case of both Mathematics and Further Mathematics, most students believed that these papers were good preparation for studying undergraduate Mathematics (see Table 5).

Table 5: Students' views of the A levels as preparation for their degree

	A level Mathematics		*AS or A level Further Mathematics*	
	N	*%*	*N*	*%*
Good preparation	558	61.9	699	76.0
Neither good preparation nor bad preparation	262	29.0	178	19.3
Bad preparation	82	9.1	43	4.7
Total	**902**	**100.0**	**920**	**100.0**

The majority of participants described A level Mathematics and Further Mathematics as good preparation for their degree. Similar proportions of participants reported this regardless of whether they had taken Further Mathematics to AS or A level (77.2% and 75.5%, respectively).

Participants' responses also appeared to be influenced by whether or not they were required to have taken Further Mathematics to be accepted onto their course. Nearly three-quarters (72.5%) of those who were not required to have taken Further Mathematics described A level Mathematics as good preparation for their degree, compared to just over half (54.7%) of those who were required to have taken Further Mathematics ($p<.001$). Similarly, those who were not required to have taken Further Mathematics were more positive about Further Mathematics as preparation for their degree ($p<.001$), with 86.5% of them describing it as good preparation compared to 68.6% of those who were required to have taken it.

Improvements to Mathematics and Further Mathematics

Participants were asked to respond to two open-ended questions: The first question, about the ways in which A level Mathematics and/or Further Mathematics could have provided better preparation for tertiary study, received 746 responses. These responses were analysed and coded using MaxQDA. The predominant themes were depth and understanding, (perceived) difficulty, content, examinations, and Applied modules. Examples from participants' comments are given to illustrate these points.

Depth and understanding: The majority of comments indicated that students would have liked more depth in both A levels. Most participants reported that A levels did not go into sufficient depth in core areas such as algebra and calculus, therefore providing insufficient preparation for undergraduate Mathematics. A smaller proportion of participants proposed that increased depth into those areas most useful for university study could be achieved by reducing the breadth of topics. For example:

Depth. It's like studying to be a pilot by only flying in a simulator.

Related to depth, although often discussed separately, was the concept of mathematical understanding. Most participants who discussed understanding in their responses indicated that the perceived lack of depth at A level hindered their understanding of mathematical concepts. They believed that this led to A level students applying particular mathematical methods with little understanding of why these methods were necessary, or the mathematical justification for doing so.

Difficulty: A substantial number of participants commented on the overall difficulty of both A levels (comments referring to the perceived difficulty of examinations were coded separately). The majority of these participants reported that A level Mathematics should be made more difficult. However, a smaller group of participants suggested that A level Mathematics was currently appropriately difficult, recognising that it is a 'service subject' for multiple degree areas and thus increasing the difficulty may reduce its accessibility.

The idea that A level Mathematics is taken by a wide range of students was reflected in participants' specific comments about Further Mathematics. Most proposed that Further Mathematics should be made more difficult, and around half of them specifically referred to the idea that students taking Further Mathematics are likely to study mathematical subjects at degree level. They proposed that Further Mathematics should cover Pure Mathematics in more depth as preparation for undergraduate Mathematics, and that the high mathematical ability of Further Mathematics candidates would enable them to cope with a higher level of difficulty.

Content: Many participants reported perceiving a lack of Pure Mathematics content in both A levels. This was the most commonly mentioned issue. The majority of these responses focused on the perceived disconnect between A level content (especially Further Mathematics) and university Mathematics. Most participants commented that there was a lack of proof and rigorous formal argument at A level, which they felt had left them poorly prepared for university study. For example:

> *It would have been very useful to start learning the mindset of a mathematician before coming to university; I've spoken to several freshers this year who are on the Maths course and all of them have said that this was the biggest and most challenging difference to school Mathematics.*

For a minority of participants, this perceived disconnect had caused concern about their choice of university course and led to a negative experience during the transition to undergraduate Mathematics:

> *More content related to topics in university – it seems like a whole different subject in university. I sometimes regret choosing it. I was so good at Maths at [A] level, I found it so easy, it came to me so fast and I loved it. At uni it's completely different and I dread going to class. I wish I had've been better prepared, [A] level does not do this!*

Examination questions: The most commonly cited overall issue (other than Pure Mathematics and proof) was the style of examination questions in both A levels. The majority of these responses described A level questions as predictable, repetitive and formulaic. This was closely related to the issue of mathematical understanding, with a similar proportion of students suggesting that it was possible to be successful in A level examinations by regurgitating known methods, without a real understanding of the actual Mathematics. The most common suggestions for improvements to rectify this perceived problem were less structural scaffolding, the use of a wider variety of contexts, and an increase in the similarity to university examinations.

Applied units: For Statistics units, the majority of participants suggested an increase in probability content and greater depth overall, which they indicated would help students' understanding of statistical theory. For Mechanics, students' opinions were split. Around one-third of participants who commented on Mechanics reported that they felt these units were too difficult and too calculation-focused, whilst approximately the same proportion perceived there to be a disconnect between Mechanics at A level and the Mechanics studied at university. Participants' comments about Decision Mathematics units reflected the negative opinions given about this strand in other data in this study.

> *The Statistics course is bad as it focuses too much on the rote application of statistical methods, which are easy to learn from scratch when their application is required, and too little on the underlying probabilistic theory and development of Statistics. Mechanics at A level feels unrelated to Mechanics at degree level, maybe the concept of calculus in Mechanics should be introduced in A level.*

Additionally, although they were in the minority of all responses, participants who mentioned Applied units in their responses often argued that both Statistics and Mechanics content should be made compulsory at A level, to ensure common grounding in both. This suggests that the introduction of compulsory applied content in the reformed A level Mathematics will be welcomed by prospective undergraduate Mathematics students.

Additional topics: There were 691 participants' responses to an open-ended question regarding additional topics that they believed should have been included in either A level Mathematics or Further Mathematics (see Table 6).

Whilst some of these topics are already covered at A level (for example, proof by induction is examined as part of Further Pure Mathematics, and the more advanced Mechanics units introduce some vector calculus), participants may not have been able to study these units at A level due to what was available for them to study at their school at the time. The most commonly suggested area for inclusion was Pure Mathematics, particularly proof, analysis, logic and group theory.

Limitations

It was not possible to state a response rate for the questionnaire because we cannot be certain how many students were contacted. There were no guarantees that (1) the questionnaire was actually sent by the departments we contacted to their students (though many departments replied to say they agreed or declined to do this), or (2) the method used to reach students was successful.

Therefore, there was potential for self-selection in terms of the departments which agreed to pass on details of the study, and in terms of the students who decided to take part once they were contacted by their department. Various factors may have played a part in certain students deciding to take part or not, including:

- Frequency with which students receive survey requests
- Time available to complete the survey
- Personal beliefs – students who felt particularly strongly either negatively or positively may have been more likely to take part
- Encouragement or presentation of the survey in communications from students' departments.

Nonetheless, a large number of students from a large number of universities took part in the study, suggesting that the methods used were effective. However, given so many of the students who took part were required to have taken Further Mathematics to be accepted onto their course, and because this study does not include responses from

Table 6: Suggested topics for inclusion at A level

	Topic	
Pure Mathematics	Proof	Proof by induction Proof by contradiction Proof by counterexample
	Analysis	
	Logic	
	Group theory	
	Number theory	
	Set theory	
Linear algebra	Matrices	Inverse matrices 3x3 matrices Operations on matrices Eigenvalues and eigenvectors Gaussian elimination
Calculus	Differentiation	Partial differentiation Higher order differential equations Second order differential equations
	Integration	Multiple integration Integration by parts
	Limits	
Mechanics	Vectors	Cross product Vector calculus 3D Vectors Vector spaces
	Kinematics	
	Circular motion	
	Quantum mechanics	
Statistics and probability	Moment generating functions	
	Expectation and variance	
	Probability theory	
Series and sequences	Fourier series	
	Convergence	
	Summation of series	
	Taylor series	
Other	Hyperbolic functions	
	Notation	
	Financial Mathematics	

students who did not take AS or A level Further Mathematics, caution must be taken when interpreting the results.

Statistical testing was conducted to ascertain whether there were any differences between students who were and were not required to have taken Further Mathematics to be accepted onto their course. This found that those who were not required to have taken Further Mathematics were often more positive about their perceptions of its usefulness in preparing them for university than those who were required to have taken it. Additionally, 90.8% of them agreed or strongly agreed that they were taught material in their first year of university that they had learned in Further Mathematics. Only 71.7% of those who were required to have taken Further Mathematics responded in the same way. These findings suggest that universities succeed in tailoring their courses to their students' mathematical backgrounds, as well as suggesting that Further Mathematics is useful preparation for undergraduate Mathematics, regardless of university entry requirements.

Further research regarding the experiences of Mathematics students who did not take Further Mathematics would be valuable. The views of students who took alternative qualifications such as the International Baccalaureate, the Cambridge Pre-U or Cambridge International A levels would also give a useful insight into students' perceptions of their mathematical preparation. Additionally, research into the mathematical backgrounds of those students who drop out of Mathematics degrees may also give an insight into the impact of students' preparation on persistence.

Discussion and conclusion

It is not currently clear how A level reform will affect the preparation of prospective undergraduate mathematicians. A shift from modular examination throughout the two years has the potential to result in a reduction in the number of students who take Further Mathematics. Until recently, many students would study four subjects in the first year of A level study, after which they would stop studying one subject and receive an AS level qualification (Gill, 2013). Without positive feedback from examination results in the first year, students may not wish to risk taking a subject they are unsure about.

The introduction of compulsory Statistics and Mechanics content in A level Mathematics will certainly be a welcome change, and will go some way in reducing the variability in students' Applied Mathematics backgrounds. However, there will be little change to the Pure Mathematics content in A level Mathematics and Further Mathematics. Topics such as matrices and complex numbers will remain in Further Mathematics rather than being moved into A level Mathematics, and there have not been any substantial changes to the proof and formal Mathematics content, both topics which new undergraduate mathematicians traditionally struggle with.

Students who participated in this study were positive about these experiences of post-compulsory Mathematics. In particular, participants valued the additional benefits of A level Further Mathematics to A level Mathematics, with 76% agreeing that it had been good preparation for their degree. Additionally, students' views of non-compulsory units suggest that Further Pure Mathematics units are by far the most beneficial in terms of the preparation they offer, and that prospective Mathematics undergraduates would benefit from a mixed background in Mechanics and Statistics, with both strands receiving a reasonable reception. Decision Mathematics appears to have had very little benefit for future undergraduate mathematicians.

Nonetheless, many students reported shortcomings in A level Mathematics and Further Mathematics, both in terms of its assessment and its content. It appears that the difficulties which students have traditionally faced with the secondary-tertiary Mathematics transition have not changed (e.g., emphasis on proof). It can only be hoped that redevelopments of the qualifications will tackle this issue, and that more students will take Further Mathematics in order to have the opportunity to study more of the useful, advanced topics prior to going to university.

Consequently, secondary teachers and careers advisers should ensure that students receive well-informed advice about useful A levels to study in preparation for certain degree subjects. Further Mathematics should

be a subject which students aim to take if they are considering pursuing the subject at university; hence it is important for schools to take advantage of initiatives such as the Further Mathematics Support Programme (2016), a government-funded project aiming to facilitate and promote the teaching of Further Mathematics in all secondary schools.

Importantly, more universities should consider either introducing Further Mathematics as a requirement for admission to their Mathematics courses or strongly recommending its study. Although universities might be reluctant to make it compulsory due to reasons relating to courses' accessibility, they should note that participants were enthusiastic about its study; most reported that they enjoyed studying it and were glad that they had done so. Prospective undergraduates should be made more aware of the views of current undergraduates regarding the usefulness of the A levels that they took at school, as well as the views of admissions tutors.

References

ACME. (2011). *Mathematical needs: mathematics in the workplace and in higher education*. London: Advisory Committee on Mathematics Education.

Anderson, J., Austin, K., Barnard, T., & Jagger, J. (1998). Do third-year mathematics undergraduates know what they are supposed to know? *International Journal of Mathematical Education in Science and Technology*, *29*(3), 401–420. Available online at: http://www.tandfonline.com/doi/abs/10.1080/0020739980290310

Baumslag, B. (2000). *Fundamentals of teaching mathematics at university level*. London: Imperial College Press.

Crawford, K., Gordon, S., Nicholas, J., & Prosser, M. (1994). Conceptions of mathematics and how it is learned: the perspectives of students entering university. *Learning and Instruction*, *4*(4), 331–345. Available online at: http://www.sciencedirect.com/science/article/pii/0959475294900051?via%3Dihub

Crawford, K., Gordon, S., Nicholas, J., & Prosser, M. (1998a). Qualitatively different experiences of learning mathematics at university. *Learning and Instruction*, *8*(5), 455–468. Available online at: http://www.sciencedirect.com/science/article/pii/S095947529800005X?via%3Dihub

Crawford, K., Gordon, S., Nicholas, J., & Prosser, M. (1998b). University mathematics students' conceptions of mathematics. *Studies in Higher Education*, *23*(1), 87–94. Available online at: http://www.tandfonline.com/doi/abs/10.1080/03075079812331380512

Darlington, E., (2014). Contrasts in mathematical challenges in A level Mathematics and Further Mathematics, and undergraduate mathematics examinations. *Teaching Mathematics and its Applications*, *33*(4), 213–229. Available online at: https://academic.oup.com/teamat/article-lookup/doi/10.1093/teamat/hru021

Darlington, E., & Bowyer, J. (2016). The mathematics needs of higher education. *Mathematics Today*, *52*(1), 9.

Edwards, P. (1996). *Implementing diagnostic testing for non-specialist mathematics courses*. London: Open Learning Foundation.

Ervynck, G. (1991). Mathematical creativity. In D. Tall (Ed.), *Advanced Mathematical Thinking* (pp.42–53). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Further Mathematics Support Programme. (2016). *Overview of FMSP*. Retrieved from http://www.furthermaths.org.uk/fmsp

Gago, J. M. (2004). *Increasing human resources for science and technology in Europe*. Brussels: European Commission.

Gill, T. (2013). *Uptake of GCE A level subjects 2012*. Cambridge: Cambridge Assessment. Available online at http://www.cambridgeassessment.org.uk/Images/109931-uptake-of-gce-a-level-subjects-2011.pdf

Gueudet, G. (2008). Investigating the secondary-tertiary transition. *Educational Studies in Mathematics*, *67*(3), 237–257.

Higher Education Statistics Agency. (2016a). *Table SN3 – Percentage of all UK domiciled entrants to full-time other undergraduate courses in 2013/14 who are no longer in HE in 2014/15*. Retrieved from https://www.hesa.ac.uk/dox/performanceIndicators/1415_B7S9/sn3_1415.xlsx

Higher Education Statistics Agency. (2016b). *Table 4 – HE student enrolments by level of study, subject area, mode of study and sex 2010/11 to 2014/15*. Retrieved from https://www.hesa.ac.uk/dox/pressOffice/sfr224/061046_student_sfr224_1415_table_4.xlsx

Hillman, J. (2014). *Mathematics after 16: the state of play, challenges and ways ahead*. London: Nuffield Foundation.

Joint Council for Qualifications. (2016). *A level Results*. Retrieved from http://www.jcq.org.uk/Download/examination-results/A levels/2016/a-ASand-aea-results

Kajander, A., & Lovric, M. (2005). Transition from secondary to tertiary mathematics: the McMaster University experience. *International Journal of Mathematical Education in Science and Technology*, *36*(2), 149–160. Available online at: http://www.tandfonline.com/doi/full/10.1080/00207340412317040

Lawson, D. (1997). What can we expect from A level mathematics students? *Teaching Mathematics and its Applications*, *16*(4), 151–157. Available online at: https://academic.oup.com/teamat/article-lookup/doi/10.1093/teamat/16.4.151

LTSN MathsTEAM. (2003). Diagnostic testing for mathematics. Retrieved from https://www.heacademy.ac.uk/system/files/diagnostic_test.pdf

Perrenet, J., & Taconis, R. (2009). Mathematical enculturation from the students' perspective: shifts in problem-solving beliefs and behaviour during the bachelor programme. *Educational Studies in Mathematics*, *71*(2), 181–198. Available online at: https://link.springer.com/article/10.1007/s10649-008-9166-9

Petocz, P., & Reid, A. (2005). Rethinking the tertiary mathematics curriculum. *Cambridge Journal of Education*, *35*(1), 89–106. Available online at: http://www.tandfonline.com/doi/abs/10.1080/0305764042000332515

Qualifications and Curriculum Authority. (2007). *Evaluation of participation in GCE Mathematics: final report*. Retrieved from http://www.ofqual.gov.uk/719.aspx

Savage, M. (2003). *Tackling the maths problem: is it far more extensive than we thought?* Paper presented at the 4th IMA Conference on the Mathematical Education of Engineering, Loughborough University.

Searle, J. (2014). *Evaluation of the Further Mathematics Support Programme*. Durham University: Centre for Evaluation and Monitoring.

Selden, A. (2012). Transitions and proof and proving at tertiary level. In G. Hanna & M. de Villiers (Eds.), *Proof and proving in mathematics education* (pp.391–420). New York: Springer.

Smith, A. (2004). *Making mathematics count*. Available online at: http://www.mathsinquiry.org.uk/report/

Smith, G., Wood, L., Coupland, M., Stephenson, B., Crawford, K., & Ball, G. (1996). Constructing mathematical examinations to assess a range of knowledge and skills. *International Journal of Mathematical Education in Science and Technology*, *27*(1), 65–77. Available online at: http://dx.doi.org/10.1080/0020739960270109

Tall, D. (1991). *Advanced mathematical thinking*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Vidal Rodeiro, C. L. (2012). *Progression from A level Mathematics to higher education*. Cambridge, UK: Cambridge Assessment.

Vidal Rodeiro, C. L., & Sutch, T. (2013). *Popularity of A level subjects among UK university students: Statistical Report Series No. 52*. Retrieved from http://www.cambridgeassessment.org.uk/Images/140668-popularity-of-A level-subjects-among-uk-university-students.pdf

Vidal Rodeiro, C. L., & Zanini, N. (2015). The role of the A* grade at A level as a predictor of university performance in the United Kingdom. *Oxford Review of*

Education, *41*(5), 647–670. Available online at: http://dx.doi.org/10.1080/03054985.2015.1090967

Williams, J. (2011). Looking back, looking forward: valuing post-compulsory mathematics education. *Research in Mathematics Education*, *13*(2), 213–221. Available online at: http://dx.doi.org/10.1080/14794802.2011.585831

Williams, J., Hernandez-Martinez, P., & Harris, D. (2010). *Diagnostic testing in mathematics as a policy and practice in the transition to higher education*. Paper presented at the conference of the British Educational Research Association, University of Warwick, Coventry.

Wolf, A. (2002). *Does education matter? Myths about education and economic growth*. London: Penguin.

Wood, L. (2001). The secondary-tertiary interface. In D. Holton (Ed.), *The teaching and learning of mathematics at university level* (pp.87–98). London: Kluwer Academic Publishers.

Question selection and volatility in schools' Mathematics GCSE results

Cara Crawford Mosaic Data Science (The study was completed when the author was based in the Research Division)

Introduction

Exam-setters face a common problem: how to condense a year or more's worth of learning into a couple of hours of test-taking. In the end, they make choices, and some topics receive more coverage in examinations than others. As a result, students may do better on one version of the test than they would do on a hypothetical alternative. In other words, for students, there is always a bit of luck involved.

But what about schools? Certainly individual students have different strengths and weaknesses within a topic area. However, there is less reason to think that the choice of test questions would have a large impact on an entire school's[1] results. Schools have recently expressed concern that test scores vary considerably from year-to-year (Headmasters' and Headmistresses' Conference [HMC], 2012), and previous research has suggested that the questions selected for a test may have small influences on candidates' grades (Benton, 2013a, 2014). If schools are not large enough to be insulated from small question-related effects on their students' grades (because each student has a non-negligible effect on the school's performance), it is possible that question-level influences on students' achievement translate to increased variability in school-level outcomes.

This research estimated the extent to which volatility in schools' scores may be attributable to changes in the selection of questions on question papers by comparing candidates' performance on two halves of the same assessment. Once student grades had been calculated for each half-test, these were aggregated within each school to form school-level outcomes for each half-test (e.g., percentage of students with a grade of C or above). Comparing the variation in schools' outcomes for their students' performance on two parts of a single test should give us some idea of the amount of variation in actual year-to-year results that could be due to changes in test questions.

Data

Data was obtained from 54,167 students who took OCR's GCSE Mathematics B (J567) qualification in the June 2014 exam session. This was chosen because it had the largest entry of any OCR GCSE and also because it consisted of a large number of questions, leaving plenty of scope for looking at variations between them. The assessment was fully linear and consisted of two written question papers. Candidates could either enter for the two Foundation Tier papers (Papers 1 and 2), covering simpler material, or for the two Higher Tier papers (Papers 3 and 4), covering upper-level material. About 56 per cent (30,310 students) were entered for the Foundation Tier (Papers 1 and 2).

All four papers had a maximum possible mark of 100, and qualification grades were based on the sum of the marks achieved on the two completed question papers. This meant that the two papers had an equal impact on final grades for the qualification.

Table 1 shows the breakdown of items (part-questions) and questions across the papers for both tiers (e.g., on Paper 1, 59 item-level marks were combined into 20 question-level marks).

Table 1: Questions and items on OCR's GCSE Mathematics B (J567), June 2014

Foundation Tier	Paper 1	59 items	20 questions
	Paper 2	65 items	23 questions
Higher Tier	Paper 3	48 items	21 questions
	Paper 4	46 items	19 questions

Methods

Overview

This research compared how the same candidates performed on two halves of a single full-length assessment. First, question papers were split by tier, with all Higher Tier questions from Papers 3 and 4 in one set and all Foundation Tier questions from Papers 1 and 2 in a second set. Within each set, questions were split into two subgroups that were as similar as possible. Candidates' marks were calculated for both subgroups of questions completed, and then mapped onto the same mark scale as the complete qualification so that grade boundaries could be set for the subgroups, and subgroup marks could be converted into grades. Each subgroup of grades in one tier was then paired with a subgroup of grades in the other tier, resulting in two combined sets of half-qualification grades. Within each school, the percentage of students achieving grades A*-C and A*-A was calculated for each half-qualification, yielding two pairs of scores for each school. Finally, school-level outcomes on the two half-qualifications were compared.

1. In this article the term 'school' is used for ease of communication instead of the more generic 'centre'. The vast majority of GCSE candidates are in schools.

Splitting questions into half-tests

Questions were split in a way that maximised the covariance between the groups, using the technique developed for calculating Guttman's λ4 reliability coefficient (Guttman, 1945). Initially questions were split into those with odd and even numbers, and then swaps between the two groups that increased the covariance were applied until no further swaps could be found. After that, the same process was repeated using additional starting splits that were assigned for the first 12 questions according to a 12 × 12 Hadamard matrix (simply a matrix that provides lots of different ways of splitting 12 questions into 2 groups so that the splits are as different as possible [Benton, 2013b]). The split yielding the highest covariance between halves (from any starting split) was retained for analysis. Benton (2013b) showed that by first splitting questions in multiple ways (e.g., even-versus-odd numbered questions, first half versus second half) and then swapping individual questions between groups to maximise the covariance between them, an optimal split can be obtained that in theory should ensure a good balance of topic areas and skills between the two halves. By maximising covariance instead of maximising correlations, one should end up with two sets of questions that have similar scales and similar distributions of scores in addition to being highly correlated.

Equating question group marks with full qualification marks

Once questions were split into two groups, equipercentile equating was used to calculate the number of marks on each question group that would correspond to each certificate-level grade on the full qualification. This was done using the *equate* package in R 3.3.1 (R Core Team, 2016) with a single-group design. The single-group design compares two tests taken by a single set of individuals (see Kolen and Brennan, 2004, for a detailed discussion of this method). This method equates scores by calculating the cumulative percentage of candidates achieving different scores on the two mark scales being compared. The score on one scale that is denoted as corresponding to a particular score on the other mark scale is chosen in a way that makes their percentile distributions (the number of candidates achieving at or below each possible score) as close to equal as possible. The intuition behind the method is that if two tests are graded to be equally difficult, then if the same students were to take both tests, the same percentage would achieve grades at or below certain points on them, and the scores that included equivalent proportions of test-takers would represent equivalent levels of performance.

Grade boundaries were then selected for the question groups based on the equated question group mark for each grade boundary on the full qualification. For example, the minimum number of marks needed to achieve a grade A* on the full (Higher Tier) qualification was 166 marks; therefore, the mark equivalent to 166 within each Higher Tier subgroup (rounded to the nearest integer) would be used as the minimum number of marks for a candidate to have achieved a hypothetical grade A* on this subgroup's questions.

Combining grades across tiers

Next, one question group from the Foundation Tier was combined with one question group on the Higher Tier so that grades across all candidates could be easily compared. Figure 1 shows how the question groups within each tier were combined into 'half-qualification' groups, with the Foundation Tier subgroups labelled as Groups W and X, and the Higher Tier subgroups denoted as Groups Y and Z.

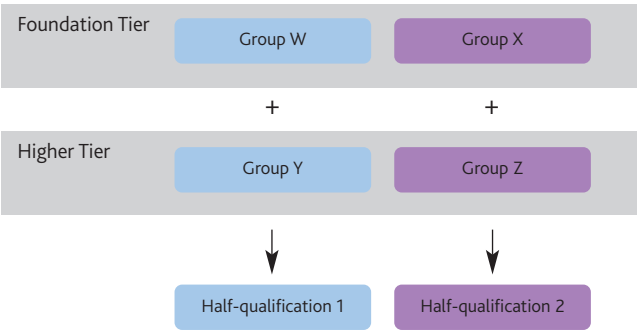

Figure 1: Combinations of question groups into half-qualifications

For each half-qualification, two school-level outcomes were computed: the percentage of students at the schools achieving grades A*-C on the half-qualification (percentage of C or above); and the percentage of students achieving grades A*-A on the half-qualification (percentage of A or above). To prevent individual students at small schools from having a disproportionately large influence on the school-level pattern of results, only schools with at least 10 students entered for the qualification were included.

Comparing schools

Correlations were computed to determine how closely a school's performance on half-qualification 1 predicted its performance on half-qualification 2. A high correlation between school outcomes on the two half-qualifications would suggest a low impact of question selection on volatility in schools' results.

Results

Table 2 examines how the questions and marks from each paper were distributed across the groups. In addition, the rightmost column shows (in bold text) the total number of questions and marks in each question group.

Table 2: Number of questions and marks in each question group by question paper

		Paper 1	*Paper 2*	*Paper 3*	*Paper 4*	*Total*
Group W	Questions	9	9			**18**
	Marks	57	40			**97**
Group X	Questions	11	14			**25**
	Marks	43	60			**103**
Group Y	Questions		8	8		**16**
	Marks		48	50		**98**
Group Z	Questions		13	11		**24**
	Marks		52	50		**102**

Table 2 shows that all question groups contained questions from more than one paper. Looking at the totals in the rightmost column of Table 2, we can see that despite differences in the number of questions in each group, they had similar numbers of marks available. This is most relevant within each tier. For example, it is good to see that even though Group Z had eight more questions than Group Y, this amounted to only four additional marks available from those questions.

Table 3: Equated scores (minimum number of marks needed to achieve each letter grade)

	Foundation Tier	*Group W*	*Group X*	*Higher Tier*	*Group Y*	*Group Z*
Range of marks	*0–200*	*0–97*	*0–103*	*0–200*	*0–98*	*0–102*
A*	-	-	-	166	82	85
A	-	-	-	133	65	68
B	-	-	-	96	46	50
C	110	53	57	59	27	31
D	91	43	48	29	12	16
E	72	33	38	14	6	7
F	54	24	29	-	-	-
G	36	16	20	-	-	-
U	0	0	0	0	0	0

Equated marks

The grade boundaries for each tier of the full qualification and the equated scores on each question group are presented in Table 3. Note that for the grades that can be obtained in both tiers (grades C, D, and E), fewer marks are needed on the Higher Tier papers than the Foundation papers. This is because the Higher Tier papers are harder, so fewer marks are needed to demonstrate the same level of mathematical knowledge.

Figure 2 compares the distribution of marks on each question group to the distribution of marks for the full qualification from which the questions were selected. The plots in the top row of Figure 2 compare the distribution of marks on the full Foundation Tier qualification against the distribution of marks in Group W (top left) and X (top right). The plots in the bottom row of Figure 2 compare the distribution of marks on the full Higher Tier qualification against the distribution of marks in Group Y (bottom left) and Z (bottom right). The main scatterplot in each figure shows the marks obtained on the

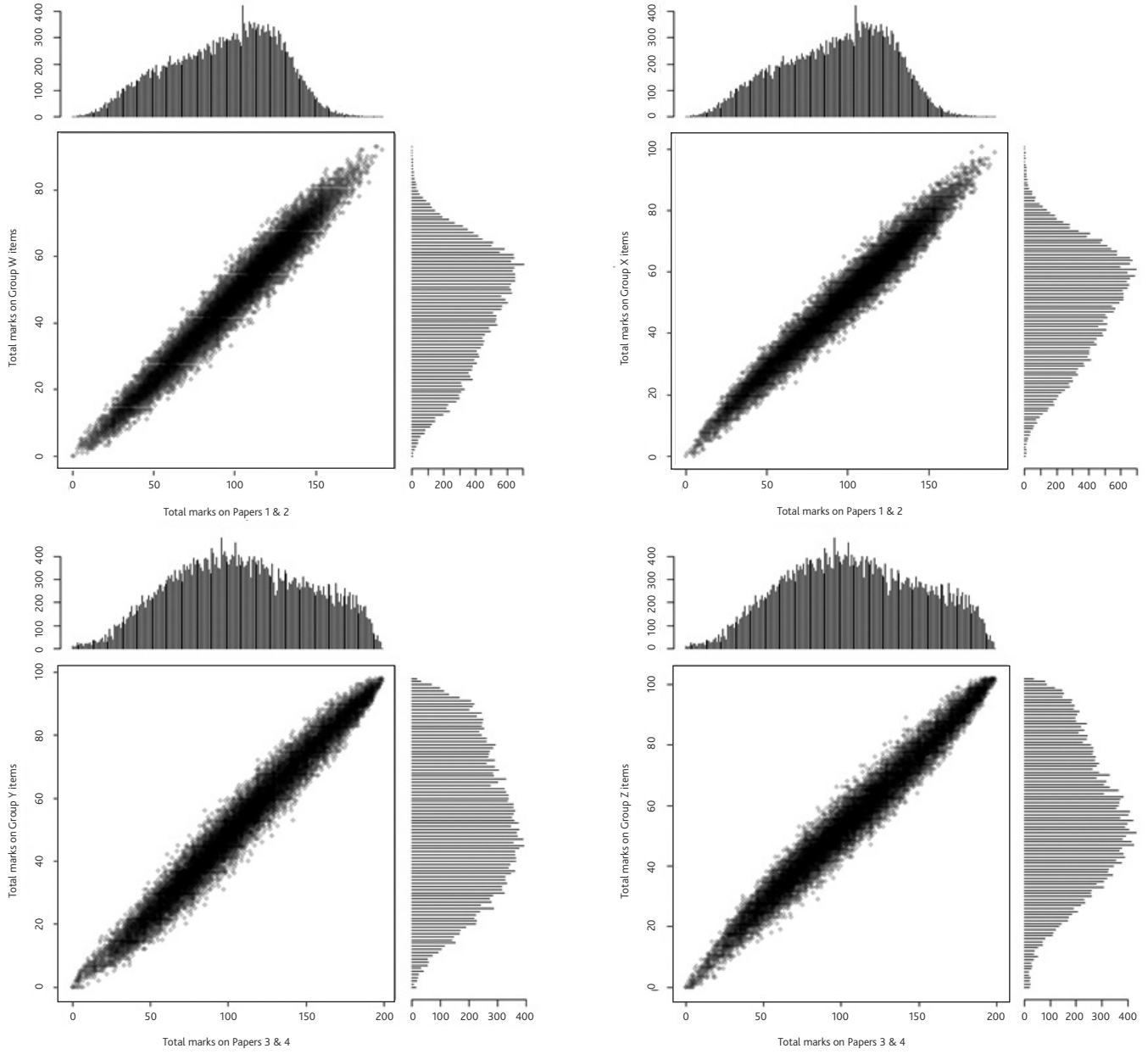

Figure 2: Marks on subgroup questions compared to total marks in each tier: Group W vs. Foundation Tier total marks (top left); Group X vs. Foundation Tier total marks (top right); Group Y vs. Higher Tier total marks (bottom left); Group Z questions vs. Higher Tier total marks (bottom right)

full papers on the x-axis and the marks obtained by the same individual on a question group on the y-axis. The fact that all four scatterplots show a positive linear relationship suggests that the question groups are all representative of the content covered in the papers they were selected from, such that higher performance on the subset of questions in each group is correlated with higher performance on the full set of questions (correlations between question groups and full qualification: Group W = .983; Group X = .982; Group Y = .986; Group Z = .986). The figures also show the distribution of marks obtained on the full qualification for each tier (above the scatterplot) and the distribution of marks on the question groups (to the right of each scatterplot). These histograms allow a comparison of the shapes of the distributions of marks between the full qualifications and the questions in each group. If the distributions have similar shapes, it suggests that a question group contains questions of a similar range of difficulty to the full qualification that its questions were selected from. Of course, these correlations are going to be positively sloped and somewhat similar because the full qualification marks include all of the marks in the question subgroups. Nonetheless, it is reassuring to see the similar patterns as these confirm that the subgroups are representative of the full set of questions.

Student half-qualification grades

A comparison of grades obtained by students on the two question groups within their tier showed that around two-thirds of students had identical grades. Specifically, 66% of Foundation Tier students had the same grade on Group W and Group X questions, and 69% of Higher Tier students had the same grade on Group Y and Group Z questions. These low-sounding levels of classification consistency[2] demonstrate how even highly correlated assessments (see top half of Table 4) can appear unreliable when analysed in this way. Although the level of absolute classification consistency does not sound particularly high, when we look at the number of grades that were either identical or just one letter grade apart (e.g., an A* and an A, or an A and a B), the figures look much better. For the Foundation Tier students, 2.4% had non-consecutive grades on the two question groups (e.g., a C on Group W's questions and an E on Group X's questions). For the Higher Tier students, the likelihood of non-consecutive grades was less than one-tenth of this size, with just 0.2% of students achieving grades on Group Y's questions that were more than one letter apart from their grade on Group Z's questions.

School-level half-qualification results

Next, half-qualifications were aggregated to school level. This resulted in two alternative sets of (half-qualification) GCSE results for each of 487 schools[3].

Correlations were computed to determine how closely a school's performance on half-qualification 1 predicted its performance on half-qualification 2. If these correlations were low, it might suggest that a good deal of school-level volatility in assessment results may be due to differences between the questions used in different exam years. However, the correlation in the percentage of grade C or above grades across schools was 0.98 and the correlation in the percentage of grade A or above grades across schools was about 0.99 (see bottom half of Table 4). In other words, looking at the variation in grade C or above results between schools, 96% of the variation in schools' half-qualification 2 results was explained by variation in half-qualification 1 results[4]. Similarly, 98% of the variation in schools' percentage of grade A or above on half-qualification 2 was explained by variation in half-qualification 1 percentages of grade A or above. This means that despite individual students sometimes receiving different scores for different groups of questions, at the school level question selection appears to have had little effect on outcomes in Mathematics.

Table 4: Correlations between half-qualification results

	Correlation coefficient between half-qualification outcomes
Student-level correlation	
Total marks (Foundation Tier)	0.944
Total marks (Higher Tier)	0.930
Grade (both tiers combined)	0.942
School-level correlations	
% grade C or above (both tiers combined)	0.978
% grade A or above (both tiers combined)	0.989

Scatterplots were created to further explore these relationships, as an overall correlation coefficient can mask variation in certain parts of a dataset. These are shown in Figure 3, with results for schools' percentage of grade A or above plotted on the left, and results for schools' percentage of grade C or above plotted on the right. On the plots, each point represents a school. Each point's position on the x-axis reflects one school's performance on half-qualification 1, and its position on the y-axis reflects the same school's performance on half-qualification 2. In both plots, a blue line shows the predicted percentage (or the most common percentage across all schools) of grade A*-A/A*-C on half-qualification 2 for each percentage of the same grades for half-qualification 1. Points are scaled by school size, with larger schools represented by larger dots on the graphs.

It appears that school size may influence differences in outcomes between the two half-qualifications, as the dots farthest from the blue lines in Figure 3 are very small (i.e., represent schools with very few students). This makes sense because a one grade change (e.g., from a C to a D) for a single student makes a larger difference in the percentage of grade C at a smaller school. To examine this potential cause of differences in half-qualification scores across schools, the absolute value of the difference in the percentage of grade A*-C and grade A*-A for half-qualification 1 versus 2 at each school was plotted against the number of students entered for the exam. These results are shown in Figure 4, with differences in the percentage of grade C or above in the plot on the left, and differences in the percentage of grade A or above in the plot on the right. The points on the graphs are semi-transparent, so a darker point indicates overlapping values for multiple schools (i.e., their points are stacked on top of each other). In both graphs, we can see that as the number of students increases, the difference in schools' achievement on the two halves diminishes. Note that the differences form lines on the graphs because results on half-qualifications differ in whole numbers of students, corresponding to a limited number of possible percentage point variations for each school.

2. The values are not low compared to those typically found in individual units or components of GCSEs and A levels – see Wheadon and Stockford (2012).
3. A total of 505 students were excluded from further analyses because they were at very small schools (entering fewer than 10 students).
4. Calculated by the fact that 0.98 (the correlation) squared is equal to 0.96.

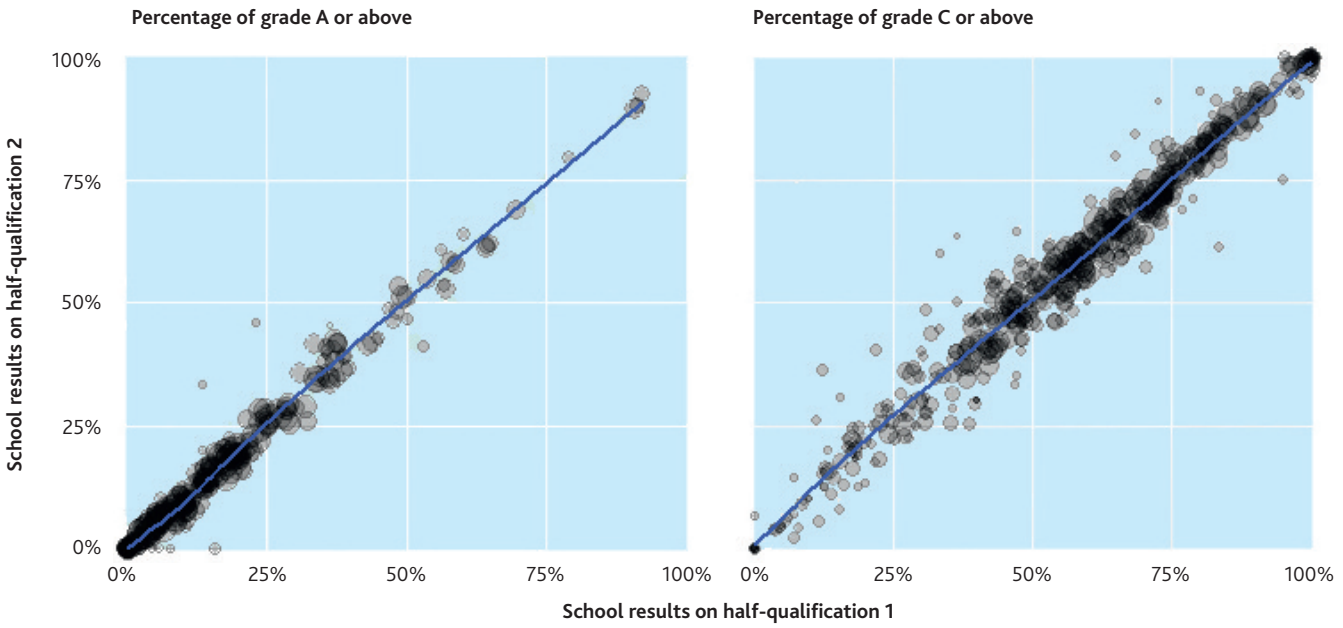

Figure 3: School-level comparison of grade A*-A (left) and grade A*-C (right) on Mathematics half-qualifications (bigger dots indicate larger schools)

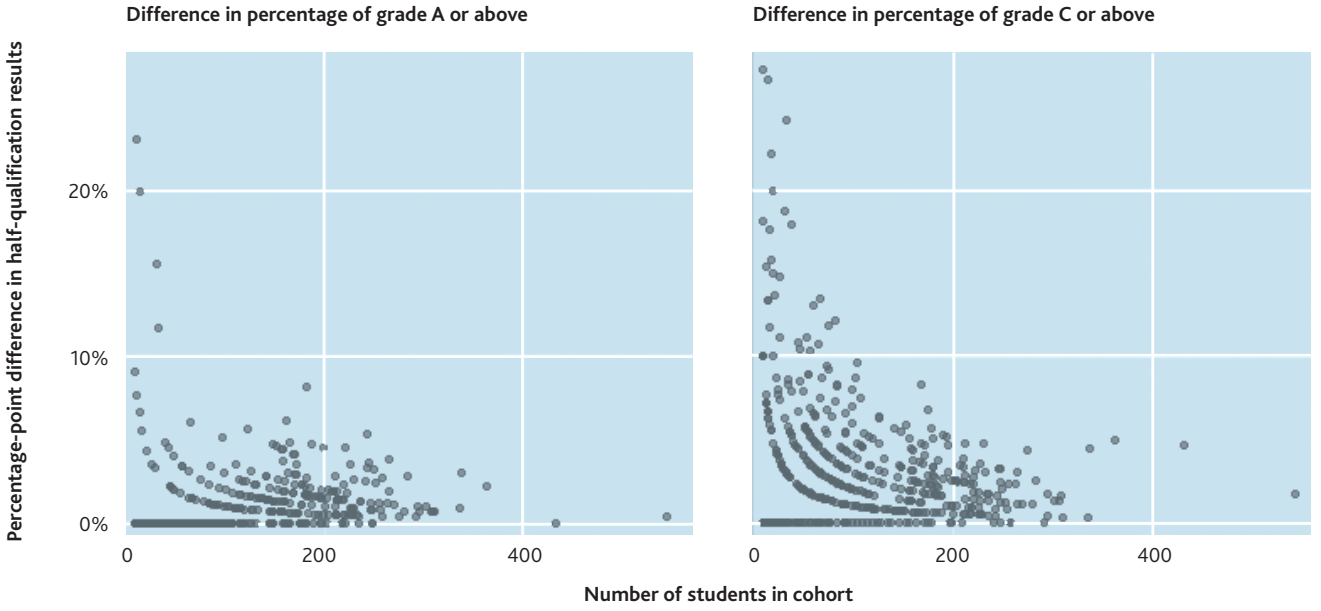

Figure 4: Absolute difference in Mathematics half-qualification results by school size

Together, these two figures indicate that as a school gets larger, volatility due to question selection decreases. To better understand this relationship, Table 5 shows descriptive statistics for the differences in percentage of grade C or above and grade A or above for schools of different sizes.

It appears that school size explains some but not all of the question-selection volatility in Mathematics results; however, even for small schools, the overall differences in performance are quite small. Because the effect of each individual question increases as the length of a test decreases, the values in Table 5 will overestimate the true amount of question-specific volatility that would occur on two full-length Mathematics qualifications. Overall, given that recent reports have considered schools to have relatively stable results when year-to-year variation is less than 10 percentage points (Ofqual, 2016), it seems that the particular selection of Mathematics questions on a given examination does not make a meaningful contribution to volatility in schools' results.

Table 5: Absolute percentage point differences in school grades on half-qualifications

	% C or above		*% A or above*		*Number of schools*
	Mean	*Max*	*Mean*	*Max*	
All schools	3.47	27.27	1.01	23.08	487
At least 50 students	2.79	13.43	1.03	8.24	356
At least 100 students	2.23	9.62	1.36	8.24	228
At least 150 students	2.04	8.33	1.53	8.24	150
At least 200 students	1.76	4.97	1.34	5.35	74

Discussion

This research investigated the potential effect of changes in questions on the same assessments in different years on volatility in schools' results over time. We did this by splitting the assessments in a single year into two shorter 'half-qualifications' and compared schools' outcomes had their students taken one of the half-qualifications instead of the entire assessment. We were interested in the extent to which schools' outcomes changed based on which of the two half-qualifications was used to determine students' grades. Our hypothesis was that if questions are comparable on two versions of an assessment (as they are supposed to be between years and as they were selected to be between halves), then students – and as a result, schools – would get similar results on both halves. Furthermore, we predicted that even if students were likely to have small differences in performance on different questions, if the two sets of questions were sufficiently alike, then these differences would not translate to differences at the school level.

The results were consistent with our predictions. For the Mathematics GCSE, it seems that little of the volatility in schools' results can be explained by differences in the questions on different versions of the tests. When students' grades were computed-based on different subsets of questions from the same question papers, the school-level outcomes were extremely similar; correlations between half-qualification percentages were extremely high, at 0.98 for the percentages of grade C or above, and 0.99 for the percentages of grade A or above.

Despite the overall pattern of results, it is not possible to determine how question selection would affect particular individual schools in particular years, other than adding an additional component of 'measurement error' to any attempts to evaluate schools based on students' test scores. Like other sources of volatility, question-selection variation will affect some schools more than others: the more students with ability levels close to the grade boundaries used to evaluate a school (e.g., borderline A/B-ability students when looking at schools' percentage of grade A and above, and C/D-ability students when looking at schools' percentage of C and above grades), the more uncertainty there will be in how that school will perform on a particular set of questions.

Caution should be used in generalising these results to other subjects. It is possible that question selection would play a larger part in the variability of schools' results in subjects that require candidates to complete fewer total questions on each question paper, or where the assessments cover a smaller range of the total taught material.

Although in general the volatility in results that occurs between exam years – and that is not explainable by differences in student ability – is quite low (Crawford & Benton, 2017), it was possible at the outset of this research that any existing volatility could be due to question selection, whereby questions on one version of an exam emphasise slightly different skills relative to another version of the same exam. Looking at question-level results for Mathematics, it appears that this explanation does not hold; for this subject (and possibly others) we must look elsewhere for explanations of volatility.

References

Benton, T. (2013a). Exploring equivalent forms reliability using a key stage 2 reading test, *Research Papers in Education*, *28*(1), 57–74. Available online at: http://dx.doi.org/10.1080/02671522.2012.754227

Benton, T. (2013b). *An empirical assessment of Guttman's Lambda 4 reliability coefficient*. Paper presented at the 78th Annual Meeting of the Psychometric Society, July 2013. Available online at: http://www.cambridgeassessment.org.uk/Images/141299-an-empirical-assessment-of-guttman-s-lambda-4-reliability-coefficient.pdf

Benton, T. (2014). Calculating the reliability of complex qualifications. *Research Matters: A Cambridge Assessment publication*, *18*, 48–52. Available online at: http://www.cambridgeassessment.org.uk/Images/174492-research-matters-18-summer-2014.pdf

Crawford, C., & Benton, T. (2017). *Volatility happens: Understanding variation in schools' GCSE results*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at: http://www.cambridgeassessment.org.uk/Images/372751-volatility-happens-understanding-variation-in-schools-gcse-results.pdf

Guttman, L. (1945). A basis for analysing test-retest reliability. *Psychometrika*, *10*, 255–282. Available online at: https://link.springer.com/article/10.1007%2FBF02288892?LI=true

HMC (2012). *England's 'examinations industry': deterioration and decay. A report from HMC on endemic problems with marking, awarding, re-marks and appeals at GCSE and A level, 2007–12*. Available online at: http://www.hmc.org.uk/wp-content/uploads/2012/09/HMC-Report-on-English-Exams-9-12-v-13.pdf

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

Ofqual. (2016). *What causes variability in school-level GCSE results year-on-year?* Ofqual/16/5956. Coventry: Ofqual. Available online at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/518409/Variability_in_Individual_Schools_and_Colleges_2016.docx_-_FINAL.pdf

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available online at: https://www.R-project.org/

Wheadon, C., & Stockford, I. (2012). Classification accuracy and consistency in GCSE and A level examinations offered by the Assessment and Qualifications Alliance (AQA) November 2008 to June 2009. In D. Opposs & Q. He (Eds.), *Ofqual's Reliability Compendium* (pp.107–139). Coventry: Office of Qualifications and Examinations Regulation.

Utilising technology in the assessment of collaboration: A critique of PISA's collaborative problem-solving tasks

Stuart Shaw Cambridge Assessment International Education and **Simon Child** Research Division

Introduction

Technological tools are increasingly becoming embedded in learning, teaching, and assessment. Advances in technology offer new opportunities for assessing collaborative learning and problem-solving in areas and contexts where assessment would otherwise not be possible. Technology can support facilitation and assessment of group collaborative learning in three ways: as a facilitator of interaction (Kreijins, Kirschner, Jochems, & Van Buuren, 2004), as a record keeper (MacDonald, 2003), or as a collaborative partner (Rosen & Tager, 2013). In this article, we briefly introduce each of these areas, before presenting an analysis of how these approaches have been enacted in relation to the Programme for International Student Assessment (PISA).

Technology as facilitator

A key issue when introducing some forms of technology relates to the concept of the 'virtual team'. Virtual teams have been described as comprising members who are geographically dispersed, but who use computer-mediated communication tools to coordinate their individual contributions (Peters & Manz, 2007). Peters and Manz (2007) argue that the higher the levels of trust between members of a virtual team, the higher the collaboration levels. This is an important consideration as members of a virtual team ordinarily have very limited face-to-face opportunities for communication in order to establish relationships. In essence, face-to-face meetings afford opportunities for members of a group to know more about each other (Mortensen & O'Leary, 2012). Unfortunately, trust takes time to grow (Henttonen & Blomqvist, 2005) and thus the role of trust in virtual teams assumes great importance (Jarvenpaa, Knoll & Leidner, 1998; Horwitz, Bravington & Silvis, 2006).

Technology as record keeper

Computer-mediated communication environments can also provide a record of activity that can be kept, replayed, and modified. The use of technology can facilitate the capturing of student activity, from which individual contributions to the collaborative process can be judged. Technology can be effectively used to provide evidence of 'artefacts' generated in the collaborative interaction in some cases, for example, by observing students as they progress on a task with a video capture software (Siemon & Scholkmann, 2015). Other examples include log files (Adejumo, Duimering, & Zhong, 2008) and capturing collaborative communication (Foltz & Martin, 2008).

Technology as collaborative partner

A challenge for assessors of collaboration is to ensure that the assessment approach can accurately capture and assess group activity and the *individual* contributions to the collaborative effort (Rosen & Foltz, 2014). Typically, the emphasis of assessment is usually at the level of individual students. This poses a challenge for the appropriate assessment of collaboration, because it is difficult to pinpoint individuals' contributions to the group effort, and to isolate these contributions from the 'noise' created by different group compositions. For example, collaborative tasks can instil conditions that encourage undesirable effects including 'social loafing' (Petty, Harkings, Williams & Latane, 1977), 'free riding' (Delton, Cosmides, Guemo, Robertson & Tooby, 2012), and competition between group members. There is also the possibility that group activities encourage a lack of different viewpoints in some cases, when group cohesion is valued over final outcomes (Rimor, Rosen & Naser, 2010).

One method of instilling a degree of standardisation in the assessment of collaboration is to utilise computer-simulated collaborative partners. The computer agent initiates collaborative behaviour, but occasionally 'tests' the test-taker by displaying some misunderstandings, or by suggesting misleading strategies. The test-taker at this point must negotiate and resolve the conflict with the computer.

Aims of the present research

There is an increasing interest in understanding how collaboration is fostered and subsequently assessed. Lee and Bonk (2014, p.10) argued that "collaborative processes and activities, as well as the technological tools for enhancing teamwork, have become increasingly critical to workplace success". The technological advancements described above, and the emerging 'assessment imperative' for so-called twenty-first century skills (Stahl, 2015) led the Organisation for Economic Cooperation and Development (OECD) to develop an assessment framework for collaboration that utilised a technological solution, which was used in PISA 2015. Each student participated in tasks where they collaborated with computer-based conversational agents only. These agents were designed to represent group members who exemplified a range of collaborative skills, knowledge, and understanding, and were programmed to introduce a degree of conflict that needed to be negotiated by the individual student. Technology was used in an attempt to control interactional boundaries, with the intention of pinpointing collaborative behaviours and traits in individual students' responses and recording them.

Our article presents the outcomes of an exercise which we conducted to map the assessment approach of PISA 2015 to pertinent facets of the collaborative process, and recent theoretical developments related to engenderment of collaboration within assessment tasks. PISA's assessment of collaborative problem-solving was mapped onto six facets of collaboration identified in a recent review of the literature (Child & Shaw, 2016, see Figure 1 for an overview) and five elements of task design that were identified in our previous review as contributing to the optimal engenderment of collaborative activity. Our present article's mapping approach afforded the opportunity to investigate in detail the

advantages and disadvantages of PISA's approach to the use of technology in its assessment of collaboration.

PISA's assessment of collaborative problem-solving using technology

The development of the PISA collaborative problem-solving assessment was built on the problem-solving framework for PISA 2012 (OECD, 2013). The OECD extended this framework to incorporate the additional concepts that focus on the collaborative aspects of problem-solving.

There are three collaborative problem-solving competencies identified in the OECD's collaborative problem-solving framework, each with a weighting:

1. Establishing and maintaining shared understanding (40–50%);
2. Taking appropriate action to solve a problem (15–30%); and
3. Establishing and maintaining group organisation (30–35%).

These competencies are synthesised with problem-solving competencies identified in PISA 2012: exploring and understanding; representing and formulating; planning and executing; and monitoring and reflecting. This synthesis is represented by a matrix of collaborative problem-solving competencies, to which specific items are designed to relate (Table 1).

Table 1: Matrix of collaborative problem-solving competencies (from OECD, 2013)

	(1) Establishing and maintaining shared understanding	*(2) Taking appropriate action to solve a problem*	*(3) Establishing and maintaining team organisation*
(A) Exploring and understanding	A1: Discovering perspectives and abilities of other team members.	A2: Discovering the type of collaborative interaction to solve the problem along with goals.	A3: Understanding roles to solve problem.
(B) Representing and formulating	B1: Building a shared representation and negotiating the meaning of a problem (common ground).	B2: Identifying and describing tasks to be completed.	B3: Describe roles and team organisation (communication protocol/rules of engagement).
(C) Planning and executing	C1: Communicating with team members about the actions to be/being performed.	C2: Enacting plans.	C3: Following rules of engagement (e.g., prompting team members to perform their tasks).
(D) Monitoring and reflecting	D1: Monitor and repair the shared understanding.	D2: Monitoring results of actions and evaluating success in solving the problem.	D3: Monitoring, providing feedback and adapting the team organisation and goals.

Part of the criteria for PISA's collaborative problem-solving construct was taking appropriate actions, using the information gathered from a previous response and an evaluation of its success (part D of the matrix). This part of the criteria was given a lower status in comparison to maintenance of the collaborative state and team organisation. The focus of PISA's 2015 assessment of collaboration was aspects of team organisation (understanding of roles within a group).

Collaborative problem-solving assessment description

The PISA assessment tasks developed to align with the OECD's collaborative problem-solving framework involved an individual student collaborating with computer-based partners, as part of a simulation of real-world collaborative activity. Each student participated in collaborative problem-solving scenarios which lasted between 5 and 20 minutes each. Within each scenario, there were several chat-based tasks where each student interacted with one or more simulated group members to solve the problem presented in the scenario. The simulated group members represented different knowledge sets and ability ranges.

There were three different task types that could reside within the overall scenarios (OECD, 2013):

1. **Consensus building:** A task type where the group needed to make a decision after considering the views, opinions, and arguments of different members.
2. **Jigsaw problems:** Each simulated group member in the task was provided with different information. The individual student had to recruit the simulated group members to pool their information and skills to achieve the group goal.
3. **Negotiations:** Group members had different amounts of information and different personal goals. Through negotiation, each student was tasked with selecting information that could be passed on so that there could be mutual win-win optimisation which satisfied overall group goals.

The assessment structure attempted to cover the 12 cells of the matrix described in Table 1, and according to the given weightings. Each item score contained within the simulation contributed to the score for only one cell of the matrix. For example, some items emphasised exploring common ground (A1), others the clarification of roles (B2), enacting plans (C2), or reflection on the successes and issues of the interaction (D3). Full-credit responses were those that targeted the maintenance of the collaborative state (Brna, 1998), whether this was achieved through refocusing on the task, offering a solution, or assigning task roles.

Construct mapping framework

Child and Shaw (2016) identified six facets related to the collaborative process (Figure 1) and suggested five prerequisites related to assessment task design and group dynamics from which collaboration can be optimally engendered. These five prerequisites are related either to the task itself or to aspects of group composition:

1. **Task should be sufficiently complex:** the problem should instil a discussion and negotiation within the group about the best course of action.
2. **Task should be ill-structured:** the task should be designed so that the appropriate course of action is not immediately outlined or discoverable.
3. **Task should only utilise non-superfluous technologies:** the task should only use technology that is essential in allowing collaboration to take place, and does not create barriers to interaction.

Figure 1: The six facets of the collaborative process (from Child & Shaw, 2016)

4. **Group member dynamics should engender negotiation:** students should be in groups where there may be differences in opinion.
5. **Group is motivated to work together:** the assessment is designed to motivate group members to work together.

The six facets of the collaborative process and the five task prerequisites were then mapped onto the example tasks of the draft collaborative problem-solving framework (OECD, 2013, 2015) described above. This helped identify which facets of collaboration were targeted within PISA 2015, and to what extent. The outcomes from the mapping exercise are outlined in Tables 2 and 3 and summarised in the following section.

Summary of mapping outcomes

The mapping exercise found that PISA 2015's tasks instilled interdependence, as the synthesis of knowledge from different group members was required for successful completion. It is questionable, however, as to whether this interdependence is 'social', because the outcome of the task is of no consequence to the computer. Johnson and Johnson (1989, 2002) argued that social interdependence is achieved when the outcome of individuals is affected by their own and others' actions, and that there is a shared overall outcome. The simulations underpinning the task in PISA 2015 mean that students cannot share an outcome with the other members of the group. However, Rosen and Foltz (2014) suggested that competitiveness may be an issue within human-to-human collaborative tasks, and thus the lack of this aspect as part of a human-computer interaction within PISA 2015 may actually improve the 'quality' of the endeavour. Furthermore, little time is needed to find common ground, which has been identified as being an important precursor to coordinate joint understandings of tasks (Clark, 1996).

Mapping also found that individual students had to share resources amongst their simulated group members to perform tasks effectively. For example, the jigsaw task was designed so that resources needed to be shared amongst group members for the task to be completed. A student was rewarded if they were able to actively share the information amongst the simulated group members and synthesise it.

One of the five prerequisite task criteria for a collaborative task was that group member dynamics should engender negotiation. In a standardised assessment of an individual student's collaboration, each student should be matched with various types of group members that will represent different collaborative skills and contexts, thus instilling discussion, negotiation, and resolution. The complexity and ill-structured nature of a task, and elements of individuals within the group, interact to afford the possibility of 'true' collaboration. The mapping for this article found that this criterion is largely met by PISA 2015. Each of the computer-based collaborative agents had their own distinct personality 'traits' which individual students had to manage and negotiate with to optimise outcomes. These were enacted periodically to test the student on particular aspects of the collaborative process.

PISA's approach to assessing collaborative problem-solving was that each set of item responses was situated within a context from which a judgement could be made about which response is optimal. For example, if an item had four potential set responses, PISA identified which response was best, relative to the other responses. There may also be a second response that received partial credit. The 'messenger' format of the interactions between the student and the computer-based collaborative partners allowed significant control in the range of situations, conflicts, and points of negotiation to which each student was exposed. In this sense, the discourse between group members was controlled to the extent that it allowed individual students to be compared on similar experiences and situations (Rosen & Tager, 2013). This approach has implications for one of the five prerequisites for a collaborative task – that the task should be ill-structured. The range of responses available to the student for each item (typically four) did not represent the full range of responses that would be possible in a real-

Table 2: Mapping of Child and Shaw's (2016) facets of collaboration to PISA 2015

Collaboration as process		
Six facets of the collaborative process (From Child & Shaw, 2016)	*Evidence from Collaborative problem-solving framework (OECD, 2013) and example tasks (OECD, 2015). Direct quotations are in italics.*	*Comment on sub-construct alignment*
1. Social interdependence	OECD (2013) states that *"Assessment items will be designed so that successful performance on the task requires collaboration and interdependency between the participants"* (p.15). This claim is supported by the 'jigsaw problem' task. It is built into the task that each group member had different information or skills. Each student needed to pool the information and recruit the skills and information from other collaborators in order to achieve the group goal. OECD (2013) states in the 'establishing and maintaining group organisation' descriptor (p.29) that *"Student acknowledges, inquires, assigns, or confirms roles taken by other group members and resources needed by other group members".* However, the 'negotiation' task implied that group members have different personal goals. This could potentially encourage some negative social interdependence if this task was misconstrued.	There are elements of both the tasks themselves and the proficiency descriptors that suggest that social interdependence was the focus of the OECD's collaborative problem-solving framework. Social interdependence is to some extent dependent on a shared outcome (Bossert, 1988), which in assessment is indicated by shared marks. This is not possible in human-computer interactions.
2. Conflict resolution	If a student attempted to move to a solution too quickly, the computer agents offer new opinions and options which required consideration and negotiation by the student.	Many of the conflicts are *implicitly* assessed. For example, there were instances where a difference of opinion or lack of focus is introduced by the computer-based partner, which the student has to negotiate.
3. Introduction of new ideas	The 'taking appropriate action' descriptor (p.29) states that *"Student takes the initiative to identify, propose, describe, or change the tasks when there are changes in the problem or when there are obstacles towards the solution".*	As each item has four response options, the student was not responsible for the creation of new ideas, but understanding when a new idea (as expressed by the response options) should be introduced into the interaction.
4. Sharing of resources	The 'establishing and maintaining shared understanding' descriptor (p.29) states that *"Student actively shares information and perspectives about self and others when it is needed".* The 'jigsaw problem' task ensured that the student and the computer-based collaborative partner/s have resources (information) that would be useful for the student to synthesise in moving towards an optimal solution.	Sharing of resources was built into both the tasks and the descriptions of performance.
5. Cooperation/ task division	Cooperativeness of group members is identified as part of 'establishing and maintaining team organisation' (p.29). The 'establishing and maintaining team organisation' level descriptor (p.29) states that *"Student's actions and communications show taking the initiative to understand and plan the different group roles that need to be taken to solve the problem."*	In the example tasks, cooperation was closely aligned with the idea of maintaining team organisation. The 'planning of group roles' may or may not involve cooperation (i.e., division of labour). .
6. Communication	Students must communicate to collaborate in the tasks. The communication stream was captured and analysed to measure the underlying processes. The 'taking appropriate action to solve the problem' descriptor (p.29) states that *"Student inquires about the actions, tasks, and plans to be completed by members of the group to solve the problem when contextually appropriate."* Students had to respond to text-based communication from computer-based collaborative partners. They had to choose from four options by clicking on the screen.	Although the scenarios were framed in a messenger-type scenario, the responses were not genuinely 'chat-like' and therefore potentially limiting. The tasks did not allow spontaneous responses. The response options offered did not reflect the full range of possible responses to each item within the scenario/s.

world context. Furthermore, it is unclear as to whether the responses available to the student were optimal, both relative to other responses, and to the infinite potential responses in the natural world. The optimum outcomes were pre-defined by the task-setter, and had in-built structure. Therefore, the degree of ill-structure might not be representative of what occurs in natural collaborative activity.

In PISA 2015, the use of computer agents meant that students had to respond to items using pre-designed textual responses. This was so that the computer agents could 'understand' the input from the student. This limited the use of other communicative strategies (for example, gestural communication) that students would potentially utilise in a human-to-human collaborative interaction, as well as the potential to share resources and introduce new ideas. Research has found that collaborators change their communication depending on their knowledge about the other communicative partners. For example, when participants were told they were interacting with a computer agent, they provided fewer references to emotion and affiliation with their partner, even when they were actually collaborating with another human (Hiyashi & Miwa, 2009). This suggests that the preconceptions that human group members had developed, based on the information that they had received previous to the commencement of the interaction, significantly influenced how they collaborated in the task.

Finally, the nature of the assessment also raises the possibility that students could be motivated to respond differently in the PISA

Table 3: Mapping of Child and Shaw's (2016) task prerequisites for collaboration to PISA 2015

Task prerequisites (From Child & Shaw, 2016)	*Evidence from PISA 2015*	*Comment on task-setting criteria alignment*
1. Task should be sufficiently complex	The collaborative task was closely tied to the concept of a 'problem'. OECD (2013) defined a problem (p.9) as existing *'when a person has a goal but does not have an immediate solution on how to achieve it'. That is, 'problem solving is the cognitive processing directed at transforming a given situation into a goal situation when no obvious method of solution is available'.*	The solutions for the tasks were unlikely to be appropriately solved by an individual student, and therefore are sufficiently complex.
2. Task should be ill-structured	Implicit in the assessment was the idea of an optimal solution/path towards a pre-defined end goal. This was a structured aspect of the assessment, as there is a final target solution that was decided by the task-setter. This goes against the conventional wisdom that task solutions should be open-ended and ill-structured.	The task was designed to have the 'appearance' of ill-structure. The student had no concept of the optimal outcome at task onset.
3. Task should only utilise non-superfluous technologies	Measurement was operationalised using computer-based agents as a means to assess collaborative skills. Students collaborated with computer-based conversational agents that represented team members with a range of skills and abilities.	The use of technology in this approach allowed a high degree of control and standardisation required for measurement.
4. Group member dynamics should engender negotiation	The group composition was determined by the task. In the examples given, there were up to two computer-based partners, each with their own characteristics. For example, one of the computer-based group members would stray off topic, and the student had to respond appropriately to keep the interaction focused.	The requirement for negotiation was built into the tasks.
5. Group is motivated to work together	Students were aware of the computer-based nature of the task, which might have affected participant motivation.	It is unclear as to how students responded to the computer-based collaborative task in terms of motivation. Motivation might be individualistic rather than shared.

collaborative tasks compared to how they would in real-life settings. This is a potential issue for one of the five task criteria – that the group is motivated to work together. Stahl (2015) suggested that the values of the collaborative framework that PISA utilised are apparent within the item choices. If this is the case, there is a potential mismatch between how a student would respond in a natural setting and how they believe they should do so to score well in the assessment. It is reasonable to assume that students were aware of the aim of the PISA assessment and the emphasis on being seen to collaborate. Whilst this issue is not unique to computer-based collaborative tasks, it does raise the question of authenticity and whether true social interdependence is possible in these tasks (Johnson & Johnson, 2002) and thus whether the individual students in PISA 2015 were motivated by different concerns compared to how they would be in a natural setting.

Conclusions and future directions

The ambition to introduce technology into the assessment of an individual's collaboration can be achieved in several ways. The challenge for assessment developers is to reconcile this ambition with considerations related to the target construct. Our article provides an analysis of the alignment of the OECD's approach to the assessment of collaboration using a previously developed theoretical framework.

PISA's assessment of collaborative problem-solving is a thorough attempt at enacting the construct of the process of collaboration, whilst using technology to provide a degree of standardisation so that comparable judgements on individual student performance can be made. Computer-based methods for the measurement of the construct of collaboration have some advantages for assessors. For example, the task can be standardised, which can facilitate the development of scoring rubrics. Furthermore, computer-based assessment can standardise aspects of interactions to facilitate the judgement of individual students. Computer-based assessment offers a 'simulation' of a collaborative task within controlled parameters. The mapping conducted in this article suggests that this approach as a means to enable 'true' collaboration, as we conceptualise it, is open to question. To illustrate this point Krkovic, Pásztor-Kovács, Gyöngyvér and Greiff (2013, p.3) suggested that "a compromise must be made and scientists have to decide if the high standardization that computers offer is worth sacrificing the face validity that human-based collaboration offers".

Our critique of PISA could lead to future work that analyses the elements of the process of collaboration that have been targeted effectively, and areas for future improvement. Specifically, it is yet to be confirmed whether a fundamental technological aspect of the assessment (the use of computer-based partners) introduces any limiting factors to the interaction. This research could provide important insights into how technology can be best utilised in the development of models of assessment for collaboration. This will be of interest to awarding organisations and others that are looking to develop qualifications in this important twenty-first century skill.

Acknowledgements

We would like to thank Sylvia Green, formerly of the Research Division, Paul Bullen-Smith, Cambridge Assessment International Education, and Helen Eccles, formerly of Cambridge International Examinations, for their insightful discussion over the course of the research, and Nick Raikes, Research Division, for his comments on an earlier draft of the article.

References

Adejumo, G., Duimering, R. P., & Zhong, Z. (2008). A balance theory approach to group problem solving. *Social Networks*, *30*(1), 83–99. Available online at: https://doi.org/10.1016/j.socnet.2007.09.001

Brna. P. (1998). Models of collaboration. *Proceedings of the Workshop on Informatics in Education, XVIII Congresso Nacional da Sociedade Brasileira de Computação*, Belo Horizonte, Brazil.

Bossert, S. T. (1988). Cooperative activities in the classroom. *Review of Research in Education*, *15*, 225–250. Available online at: http://journals.sagepub.com/doi/abs/10.3102/0091732X015001225

Clark, H. H. (1996). *Using language*. New York: Cambridge University Press.

Child, S. F. J., & Shaw, S.D. (2016). Collaboration in the 21st century: Implications for assessment. *Research Matters: A Cambridge Assessment publication*, *22*, 17–22. Available online at: http://www.cambridgeassessment.org.uk/Images/374626-collaboration-in-the-20th-century-implications-for-assessment.pdf

Delton, A. W., Cosmides, L., Guemo, M., Robertson, T. E., & Tooby, J. (2012). The psychosemantics of free riding: Dissecting the architecture of a moral concept. *Journal of Personality and Social Psychology*, *102*(6), 1252–1270. Available online at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3365621/

Fawcett, L. M., & Garton, A. F. (2005). The effect of peer collaboration on children's problem-solving ability. *The British Journal of Educational Psychology*, *75*(2), 157–169. Available online at: http://onlinelibrary.wiley.com/doi/10.1348/000709904X23411/fulll

Foltz, P. W., & Martin, M. J. (2008). Automated communication analysis of teams. In E. Salas., G. F. Goodwin., & S. Burke (Eds.). *Team effectiveness in complex organizations and systems: Cross-disciplinary perspectives and approaches* (pp.411–431). New York: Routledge.

Henttonen, K., & Blomqvist, K. (2005). Managing distance in a global virtual team: The evolution of trust through technology-mediated relational communication. *Strategic Change*, *14*(2), 107–119. Available online at: http://onlinelibrary.wiley.com/doi/10.1002/jsc.714/full

Hiyashi, Y., & Miwa, K. (2009). Cognitive and emotional characteristics of communication in human-human/human-agent interaction. In J.E. Jacko (Ed.). *Human-Computer interaction: Ambient, ubiquitous and intelligent interaction* (pp.267–274). Berlin: Springer-Verlag.

Horwitz, F. M., Bravington, D., & Silvis, U. 2006. The promise of virtual teams: Identifying key factors in effectiveness and failure. *Journal of European Industrial Training*, *30*(6), 472–494. Available online at:https://link.springer.com/chapter/10.1007%2F978-3-642-02580-8_29?LI=true

Jarvenpaa, S. L., Knoll, K., & Leidner, D. (1998), 'Is Anybody Out There? Antecedents of trust in global virtual teams. *Journal of Management Information Systems*, *14*(4), 29–64.

Johnson, D. W., & Johnson, R. T. (1989). *Cooperation and competition: Theory and research*. Edina, MN: Interaction Book Company.

Johnson, D. W., & Johnson, R. T. (2002). Learning together and alone: Overview and meta-analysis. *Asia Pacific Journal of Education*, *22*(1), 95–105. Available online at: http://www.tandfonline.com/doi/pdf/10.1080/0218879020220110

Johnson, D. W., Johnson, R. T., & Smith, K. (2007). The state of cooperative learning in postsecondary settings. *Educational Psychology Review*, *19*(1), 15–29. Available online at: https://link.springer.com/article/10.1007/s10648-006-9038-8

Kreijins, K., Kirschner, P. A., Jochems, W., & Van Buuren, M. A. (2004). Determining sociability, social space, and social presence in (a)synchronous collaborative groups. *CyberPsychology & Behaviour*, *7*(2), 155–172. Available online at: https://doi.org/10.1089/109493104323024429

Krkovic, K., Pásztor-Kovács, A., Gyöngyvér, M. & Greiff, S. (2013). New technologies in psychological assessment: The example of computer based collaborative problem solving assessment. In D. Whitelock., W. Warburton., G. Wills., & L. Gilbert (Eds.). *Conference proceedings of the CAA 2013 International Conference*, University of Southampton.

Lai, E. R., & Viering, M. (2012). Assessing 21st century skills: Integrating research findings. *Paper presented at the annual meeting of the National Council on Measurement in Education*, Vancouver, B.C., Canada.

Lee, H., & Bonk, C. (2014). Collaborative learning in the workplace: Practical issues and concerns. *International Journal: Advanced Corporate Learning*, *7*(2), 10–17. Retrieved from http://dx.doi.org/10.3991/ijac.v7i2.3850

MacDonald, J. (2003). Assessing online collaborative learning: process and product. *Computers & Education*, *40*(4), 377–391. Available online at: http://www.sciencedirect.com/science/article/pii/S0360131502001689

Mortensen, M. & O'Leary, M. (2012). *Managing a virtual team*. Retrieved 18 February, 2015, from http://blogs.hbr.org/cs/2012/04/how_to_manage_a_virtual_team.html

OECD (2013). PISA 2015: Collaborative problem-solving framework. Available online at: http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

OECD (2015). PISA 2015: released field trial cognitive items. Available online at: http://www.oecd.org/pisa/pisaproducts/PISA2015-Released-FT-Cognitive-Items.pdf

Peters, L. M., & Manz, C. C. (2007). Identifying antecedents of virtual team collaboration. *Team Performance Management: An International Journal*, *13*(3/4), 117–129. Available online at: http://www.emeraldinsight.com/doi/abs/10.1108/13527590710759865

Petty, R. E., Harkins, S. G., Williams, K. D., & Latane, B. (1977). The Effects of Group Size on Cognitive Effort and Evaluation. *Personality and Social Psychology Bulletin*, *3*(4), 579–582. Available online at: http://journals.sagepub.com/doi/abs/10.1177/014616727700300406

Rimor, R., Rosen, Y., & Naser, K. (2010). Complexity of social interactions in collaborative learning: The case of online database environment. *Interdisciplinary Journal of E-Learning and Learning Objects*, *6*(1), 355–365.

Rosen, Y. (2014). Comparability of conflict opportunities in human-to-human and human-to-agent online collaborative problem solving, *Tech Know Learn*, *19*, 147–164. Available online at:https://link.springer.com/article/10.1007/s10758-014-9229-1

Rosen, Y., & Foltz, P. W. (2014). Assessing collaborative problem solving through automated technologies. *Research and Practice in Technology Enhanced Learning*, *9*(3), 389–410. Available online at: http://meyda.education.gov.il/files/Scientist/RosenFoltz2014.pdf

Rosen, Y., & Tager, M. (2013). *Computer-based assessment of collaborative problem solving skills: Human-to-agent versus human-to-human approach*. Research & Innovation Network. Pearson Education.

Siemon, J., Scholkmann, A., & Bloom, K-D. (2015). 'Time on task' in collaborative learning: Influence of learning goal motivation and group composition. *Paper presented at the European Conference for Education Research*, Budapest, 7th–11th September, 2015.

Stahl, G. (2015). PISA 2015: *Assessing collaborative learning*. Available online at: http://gerrystahl.net/pub/cscl2013pisa.ppt.pdf

Webb, N. M. (1991). Task-related verbal interaction and mathematical learning in small groups. *Research in Mathematics Education*, *22*(5), 366–389. Available online at: http://www.jstor.org/stable/749186

Partial absences in GCSE and AS/A level examinations

Carmen Vidal Rodeiro Research Division

Introduction

There are certain situations in which a candidate does not have a mark for a unit/component in a GCSE or AS/A level examination. For example, if they were ill on the day of the exam, if their paper was lost (e.g., at the centre (school), in the post, at the scanning bureau or at the awarding body's offices) or if their controlled assessment was invalid as a result of individual or centre malpractice.

Subject to certain rules, the awarding body can calculate an estimated mark for the unit/component with the missing mark to enable the candidate to certificate, rather than having to wait for the next assessment opportunity. The conditions under which an estimated mark can be awarded are set out by the Joint Council for Qualifications (JCQ, 2016).

There have been reports in the press (e.g., Espinoza, 2015; Linning, 2015) about awarding bodies 'guesstimating' hundreds of students' grades. However, a spokesman from the Office of Qualifications and Examinations Regulation (Ofqual) (quoted in the above press reports) said that a very small number of marks can be and are estimated each year and only in some very specific circumstances. In fact, he said, this number represents just a very small fraction of the number of overall papers marked.

The aims of the research described here were as follows:

1. To investigate the numbers of unit/component marks in GCSE and AS/A level qualifications awarded by the OCR awarding body that were estimated in a specific session.
2. To evaluate current and potential new method(s) for estimating missing marks. In particular, this research explored the use of statistical methods for handling missing data, specifically *regression imputation*, to estimate the mark for a missing unit/component in GCSE and AS/A level qualifications. The marks (and grades) obtained in this way were compared with the marks (and grades) obtained applying two different methods currently used by some of the awarding boards in England: the *z-score method* and the *percentile (cum% position) method*.

Data and methodology

In this research, unit/component level data from the OCR awarding body (June 2015 session) was used.

For the investigation of methods for estimating missing marks, the following analyses were carried out:

- **Simulation of missing data:** missing marks for a specified number of candidates were simulated in several GCSE and AS/A level units/components. Different strategies, which are described later, were used for this.
- **Estimation of the missing marks using three different methods:** Regression imputation, *z-score method* and *percentile method*.

Creation of missing marks

Several OCR qualifications, both at GCSE and AS/A level, with different structures (e.g., different number of units; different types of assessment) were selected for analysis. Table 1 (on page 24) gives details of the specifications included in this work.

In each of the specifications listed in Table 1, units were selected as shown in Table 2 (on page 24), and missing marks were then generated.

Partial absences for candidates certificating in June 2015 in OCR qualifications (GCSE and AS/A levels) were examined to give an idea of the numbers of candidates who are issued estimated grades in a given session. There were 19 GCSE units/components with at least 40 candidates with missing marks, and 11 units with at least 60. At AS/A level, there were several units with more than 40 candidates who had estimated marks but just 1 with more than 50. Taking this information into consideration, it seemed reasonable to select, in each unit/component listed in Table 2, 60 candidates to create missing marks for.

The different strategies to create the missing marks were as follows:

1. Candidates were selected at random and their marks in the unit/component of interest were set to missing.
2. The probability of having an absent mark for the unit/component of interest was modelled, using a logistic regression, as a function of the overall qualification grade:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Grade_i + e_i$$

where p_i is the probability of candidate i being absent and *Grade* is the overall qualification grade.

This probability was used to ensure that, if certain grades were more prominent amongst candidates with partial absences, this was reflected in the sample.

3. Using unit/component level data from OCR qualifications, candidates with missing marks in June 2015 (due to missing scripts or special consideration) in any unit/component were selected. This was done for GCSE and AS/A level units/components separately.

Candidates in the unit/component of interest with a missing mark in any other unit/component (in a qualification of the same level) had their mark set to missing. If there were more than 60 candidates fulfilling this condition, 60 were selected at random amongst them.

Methods to estimate missing marks

Z-score method

In order to estimate a missing mark when a candidate is absent from an examination in a specification which uses uniform marks (most unreformed GCSE and AS/A levels), most of the JCQ awarding bodies have been employing the same procedure, known as the *z-score method*.

Table 1: OCR qualifications considered in the analyses, June 2015

Qualification	*Subject*	*OCR specification*	*Unit/Component*	*Type of assessment*	*Weighting*	*Maximum UMS[a] marks*
AS	Biology	H021	F211	written paper	30%	90
			F212	written paper	50%	150
			F213	coursework	20%	60
AS	Media Studies	H140[b]	G321	coursework	50%	100
			G322	written paper	50%	100
			G323	written paper	50%	100
GCSE	Business and Communication Systems	J230	A265	written paper	50%	120
			A266	controlled assessment	25%	60
			A267	practical examination	25%	60
GCSE	Religious Studies B	J621	B601	written paper	25%	50
			B602	written paper	25%	50
			B603	written paper	25%	50
			B604	written paper	25%	50
GCSE	Mathematics B (linear)	J567[c]	01 (F)	written paper	50%	100
			02 (F)	written paper	50%	100
			03 (H)	written paper	50%	100
			04 (H)	written paper	50%	100

a. Uniform Mark Scale used for modular assessments. For the Mathematics B (linear) J567, this column shows the maximum possible raw mark.
b. To obtain an AS level in Media Studies candidates needed to take unit G321 and either unit G322 or unit G323.
c. To obtain a GCSE in Mathematics B candidates needed to take components 01 and 02 (Foundation Tier [F]) or 03 and 04 (Higher Tier [H]).

Table 2: Units/Components for which missing marks were generated, June 2015

Unit/ Component	*Qualification/Subject*	*Type*
F212	AS Biology	written paper
F213	AS Biology	coursework
G322	AS Media Studies	written paper
A267	GCSE Business and Communication Systems	practical examination
B601	GCSE Religious Studies B	written paper
J567/01	GCSE Mathematics B (linear)	written paper

Under this procedure, the difference between the candidate's estimated mark and the performance of candidates generally on the unit in question is the same as the average difference between the candidate's performance and the performance of candidates generally on the other units.

If the candidate performed on average slightly better (or worse) than candidates generally in other units, then the estimate for the missing mark will be slightly above (or below) the general performance on the unit of interest. The difference between the performance of the candidate in question and the performance of candidates generally is measured in terms of standard deviations. The number of standard deviations above or below the mean is called the *z-score*.

An illustration of how the method works is given below.

Example

In a three-unit specification, the average uniform marks and the standard deviation for all candidates on Units 1, 2 and 3 are given in Table 3 below. Unit 1 accounts for 30% of the assessment, Unit 2 for 50% and Unit 3 for 20%.

Let us assume a candidate scores 58 on Unit 1 and 104 on Unit 2, but is absent for Unit 3. The estimated mark for Unit 3 is calculated as shown below.

Table 3: Candidate's and average performance

	Weighting	*Average uniform marks*	*Standard deviation*	*Candidate's mark*
Unit 1	30%	50	8	58
Unit 2	50%	80	12	104
Unit 3	20%	38	3	*absent*

The candidate's mark on Unit 1 is one standard deviation (8 marks) above the average for all candidates in that unit. The candidate's mark on Unit 2 is two standard deviations (24 marks) above the average for that unit. So, taking into account the weightings of the units, the average of the standard deviations is:

$$\frac{30 \times 1 + 50 \times 2}{30 + 50} = 1.625$$

Thus, the estimated mark for Unit 3 is:

$$\text{average mark} + 1.625 \times \text{standard deviation} = 38 + 1.625 \times 3 = 42.875,$$

which is rounded to 43.

An alternative to the above method is to calculate the z-score based on the aggregation of the marks in the available units/components (candidate's total score) rather than at unit/component level. This method would have the advantage of taking into account the correlation between the marks in the different units. However, this might not be feasible in practice, as in some qualifications candidates take different optional units/components.

Percentile (cum% position) method

The principle for this method is to identify unit(s)/component(s) from the same specification for which the candidate has marks, and to award the candidate a mark for the missing unit that, as nearly as possible, places them at the same percentile of the cohort as they have achieved on the unit(s)/component(s) being used in the calculation. Where the relevant percentile occurs between two possible marks, the higher mark is awarded.

It might be recommended to align the candidate's mark in the missing unit/component with fewer units/components if there is a good reason for discounting a unit/component. For example, if a coursework unit/component is being used to estimate the mark in a written unit/component and performance is not expected to correlate much between those two units/components, it might make sense to exclude it.

An illustration of how the method works is given below.

Example

Let us assume that in a three-unit specification, a candidate is missing the mark for Unit 2, and this will be calculated using the candidate's performance in Units 1 and 3 (Table 4 below). Table 5 shows extracts of the cumulative mark distributions for Units 1 and 3.

Table 4: Candidate's performance

	Weighting	*Candidate's mark*
Unit 1	30%	72
Unit 2	50%	*absent*
Unit 3	20%	55

Table 5: Mark distributions, Units 1 and 3

	Mark on unit	*Cumulative percentage of candidates*
Unit 1	73	18.37
	72	22.12
	71	22.12
Unit 3	56	3.49
	55	7.42
	54	7.95

The cumulative percentages that correspond to marks 72 and 55 in Units 1 and 3 are 22.12 and 7.42 respectively. The next step in the method is to take the average of these two figures, taking into account the weights of the units:

$$\frac{22.12 \times 30 + 7.42 \times 20}{30 + 20} = \frac{663.6 + 148.4}{50} = \frac{812.0}{50} = 16.24$$

Looking through the mark distribution for the unit with the absent mark (Unit 2), displayed in Table 6 below, we find out the mark that corresponds to that cumulative percentage. The nearest marks on Unit 2 to one that gives 16.24 per cent of candidates are 123 and 122. A mark of 123 is hence taken.

Table 6: Mark distribution, Unit 2

Cumulative percentage of candidates	*Mark on Unit 2*
13.29	124
15.73	123
17.65	122

Regression imputation

Many missing data methods fall under the general heading of imputation. The basic idea of those methods is to substitute each missing value with some reasonable prediction (imputation). There are lots of different ways to impute missing values. In this research, *regression imputation* was used.

A regression model was fitted to predict the values of a dependent variable (marks in the unit/component of interest) based on other independent variables potentially related to the missing data (e.g., performance in other units/components of the same qualification, characteristics of the candidates). The model was then used to impute values in cases where the dependent variable was missing.

Two different regression techniques were used: *ordinary least squares* (OLS) and *quantile regression*. OLS models the conditional mean of the response or dependent variable as a function of one or more independent variables. Quantile regression models the conditional quantiles (in particular, the median), rather than the mean.

In this research, the following information for each candidate was available:

- Performance in other units/components for the same qualification.
- A measure of overall performance. This was calculated, using principal component analysis, for each candidate who had taken at least one OCR assessment in the June 2015 session. It reflects the marks achieved on all the assessments taken (excluding the score for the particular unit/component being imputed).
- Characteristics of the candidate (gender, year group, socio-economic level[1]).
- Characteristics of the school (type[2], overall attainment of its pupils[3]).

Some of the information on the candidates was obtained from the OCR awarding body and some from the National Pupil Database (NPD)[4]. Information from the NPD (e.g., socio-economic level, type of school and the overall attainment in each school) was matched to OCR data using candidate and centre numbers.

An illustration of how the method works is given below.

Example

Let us assume that a candidate sat a three-unit specification and was missing the mark in Unit 2 (as seen in Table 4 above).

A very simple linear regression model, just to illustrate the method, was fitted. The dependent variable was the mark in Unit 2 and the independent variables the marks in Unit 1 and Unit 3 (more complex models will be used in the analyses presented in this article). The fitted model is as follows:

1. The socio-economic level was measured by the IDACI (Income Deprivation Affecting Children Index). This index measures the percentage of children in a small area around the student's home who live in families that are income deprived.
2. Independent schools, selective schools, state-maintained schools (including comprehensive, secondary modern and academy schools), sixth form colleges and further education (FE) colleges.
3. School average GCSE or A level performance, depending on the qualification analysed.
4. The NPD, compiled by the Department for Education, is a longitudinal database for all children in schools in England, linking student characteristics to school and college learning aims and attainment. The NPD holds pupil and school characteristics such as gender, ethnicity or level of deprivation (IDACI) matched to pupil level attainment data.

$$\widehat{Unit\ 2} = 5.93 + 1.21\ (Unit\ 1) + 0.44\ (Unit\ 3)$$

The model above is used to predict the mark in Unit 2 for the candidate shown in Table 4:

$$5.93 + 1.21\ (72) + 0.44\ (55) = 117.25$$

The estimated mark, based on the candidate's performance in the other two units that contribute to the qualification, would be 117.

Comparison of methods

In order to compare the performance of the three methods described above, the following measures were used:

1. An indicator of how close the estimated mark was to the actual mark, over all candidates with estimated marks (n):

   $$\sum_{i=1}^{n} abs\ (mark - estimated\ mark)$$

   Quantile regression is designed to be the line of best fit that minimises this indicator.

2. Correlation coefficients between estimated and actual marks:

   $$Corr\ (marks, estimated\ marks)$$

3. The root mean square error (RMSE) of the estimated marks:

   $$\sqrt{\frac{\sum_{i=1}^{n} (mark - estimated\ mark)^2}{n}}$$

   where the sum is over all candidates with estimated marks (n).

   This measure is minimised by OLS regression.

These three statistics were used to compare the performance of the three methods for estimating missing marks using the candidates' marks. Qualification grades awarded based on estimated marks were also compared with actual grades, using a variety of simple descriptive statistics.

Results

Partial absences in GCSE and AS/A level units

A brief investigation into partial absences for candidates certificating in June 2015 was carried out to illustrate the numbers of candidates who were issued estimated grades. In June 2015, very small numbers of candidates received an estimated grade. The AS/A level unit with the highest number of estimated marks was F213 (Practical Skills in Biology 1), where 53 candidates out of 36,582 (0.14% of the entry) were missing the mark. At GCSE, B712 (Science modules B2, C2 and P2) had 112 candidates with an estimated mark (0.23% of the entry).

Overall, only 1,073 AS/A level missing scripts had estimated marks in June 2015, which is below 0.1% of the total number of AS/A level scripts marked by OCR. Similarly, at GCSE, 2,289 (0.08%) missing scripts had estimated marks.

In this research, instances of malpractice were not considered and only absences due to missing scripts or special consideration (the candidate was ill on the day of the exam) are included in the tables.

Estimating missing marks

This section reports on the results of the three different methods used to estimate the generated missing marks in the units/components listed in Table 2.

In the case of the regression imputation (for both the OLS and the quantile regression techniques), two different models were fitted:

(a) A model only including, as independent variables, the marks for the other units/components in the specification. In this case, the information used in the imputations is the same as the information included in the percentile and z-score methods.

Table 7: Differences between actual and estimated marks

Unit/ Component	*Missing data generation*	*Sum of absolute differences*					
		Percentile	*z-score*	*Regression imputation*			
				OLS		*Quantile regression*	
				(a)	*(b)*	*(a)*	*(b)*
F212	Scenario 1	990	964	823	746	813	742
	Scenario 2	986	963	844	661	838	659
	Scenario 3	906	903	770	613	761	608
F213	Scenario 1	420	414	349	321	349	321
	Scenario 2	482	506	386	337	378	332
	Scenario 3	463	473	485	347	482	342
G322	Scenario 1	823	810	678	634	676	629
	Scenario 2	780	752	551	561	549	562
	Scenario 3	-[a]	-	-	-	-	-
A267	Scenario 1	401	393	417	384	413	385
	Scenario 2	461	432	426	391	423	394
	Scenario 3	-	-	-	-	-	-
B601	Scenario 1	323	319	305	300	310	303
	Scenario 2	273	260	262	239	261	238
	Scenario 3	277	273	280	270	278	268
J567/01	Scenario 1	330	329	309	283	310	290
	Scenario 2	386	404	401	371	401	371
	Scenario 3	390	398	380	359	382	363

a. The '-' in the table indicates that there were no candidates taking the unit who had partial absences in any other unit at this level (GCSE or AS/A level) in the June 2015 session.

(b) A model including, as independent variables, the marks for the other units/components in the specification, information about the candidates (gender, year group), their overall performance and characteristics of the schools.

Note that model (b) has been proposed for research purposes and that there are practical limitations (e.g., data acquisition, timescales) which might mitigate against its use in an operational setting.

Table 7 shows the differences between the actual and estimated marks (absolute values) for all units/components considered in this research and in all three scenarios described in the Data and methodology section. This statistic indicates how close the estimated marks are to the actual marks (over all partial absent candidates) and, therefore, the method with the smallest value would be more desirable.

OLS and quantile regression provided very similar results for all units/components in the three different scenarios. Furthermore, and not unexpectedly, the models including additional information on the candidates and the schools provided smaller values for the differences between actual and estimated marks. Therefore, from this point onwards, the focus is on the results from the OLS model (b). These are compared with the results obtained using the percentile and z-score methods.

For all units considered and regardless of the different mechanisms used to create the partial absence data (scenarios 1 to 3), the estimated marks calculated via regression imputation provided the best result in terms of the sum of absolute differences between actual and estimated marks.

The percentile and z-score methods provided very similar results. However, it should be noted that the differences between actual and estimated marks were usually smaller when using the z-score method than when using the percentile method (the z-score method yielded smaller total errors in 12 out of 16 tests above). Only for Unit F213 and Component J567/01, and in scenarios 2 and 3, did the percentile method seem to perform slightly better.

Due to the fact that the maximum UMS marks available was not the same for all units (see Table 1), it was not possible to compare the values for the sum of absolute differences across all units in Table 7. However, G322 and J567/01 had the same maximum UMS marks (100 marks) and comparisons were therefore possible. In this case, the measure of discrepancy was smaller in all scenarios and for all methods (particularly in the z-score and percentile methods) in Component J567/01. The marks in this component were estimated using information from Component J567/02, which assesses similar content in the same way: that is, via an external written paper. In contrast, the marks for Unit G322, a written paper, were estimated using information from a coursework unit (G321).

Therefore, when the marks in the units involved in the analysis do not correlate strongly (which is usually the case between coursework and written paper units), neither the z-score method nor the percentile method seems appropriate for estimating the missing marks. The regression imputation method, which includes further information about the candidates, their overall performance and takes into account characteristics of the schools, provides better estimates for the missing marks.

It should be noted that the method itself, rather than the additional information, may be what makes regression imputation a better option for estimating missing marks. However, if we look at the results from OLS model (a), which was based on the same information as the percentile and z-score methods, there were improvements in some cases (particularly F212, F213 and G322) but not in all, and the differences between actual and estimated marks were sometimes much smaller in OLS model (b) than in OLS model (a). This shows that there might be an effect of method but often the effect of the additional data is bigger.

In order to investigate further the effect of using the mark in a coursework unit to estimate the mark in a written paper, missing marks in Unit F212 were estimated leaving out the marks in Unit F213. Table 8 below shows the results.

By comparing the first three rows of Table 7 with the same rows in Table 8, we can see that in the percentile and z-score methods, the differences between actual and estimated marks are smaller (overall) when the coursework marks are ignored. Results for the regression imputation (OLS model [b]) are very similar in both tables.

Therefore, before estimating any missing marks, it seems worthwhile to decide which unit(s)/component(s) of the same specification should be considered and which ones should be discounted. This seems fairly relevant when using the percentile or z-score method, but not so much when using regression imputation.

Table 8: Differences between actual and estimated marks (F212 based on F211)

Unit/ Component	*Missing data generation*	*Sum of absolute differences*		
		Percentile	*z-score*	*Regression imputation OLS (b)*
F212	Scenario 1	842	829	747
	Scenario 2	745	728	666
	Scenario 3	737	730	612

The RMSE statistic was also calculated for all units/components and in all scenarios. The results were consistent with those presented in Table 7, that is, the marks calculated via regression imputation provided the closest estimates (overall) and the percentile and z-score methods provided very similar results.

Correlations between actual marks and marks estimated by the three proposed methods are given in Table 9 below. In this case, the method with the highest correlations would be the best to use.

In all scenarios, correlations were highest when the missing marks were estimated by regression imputation, particularly when including additional data in the models OLS model (b) and quantile regression model (b). The differences between regression imputation and the percentile and z-score methods were bigger when different types of units were involved in the analyses (e.g., Units F212, F213, A267 and, particularly, G322). However, for Unit B601 and Component J567/01 (marks in a written paper were estimated using performance in written papers only), correlations were fairly high, and similarly, independent of the estimation method.

Grading based on estimated marks

In this section, qualification grades based on marks calculated by the three different methods discussed earlier are presented and compared with actual grades.

Firstly, the percentages of candidates who achieved the same estimated grade as the actual grade by each of the three methods considered in this work are displayed in Table 10. This shows that, as was reported for the marks in the previous section, the regression imputation method (either using OLS or quantile regression) provides, overall, the most accurate results.

Table 9: Correlations between actual and estimated marks

Unit/ Component	*Missing data generation*	*Correlations*					
		Percentile	*z-score*	*Regression imputation*			
				OLS		*Quantile regression*	
				(a)	*(b)*	*(a)*	*(b)*
F212	Scenario 1	0.794	0.815	0.864	0.881	0.864	0.880
	Scenario 2	0.627	0.672	0.769	0.840	0.770	0.840
	Scenario 3	0.785	0.778	0.845	0.882	0.845	0.881
F213	Scenario 1	0.516	0.547	0.546	0.569	0.546	0.566
	Scenario 2	0.221	0.199	0.200	0.296	0.200	0.297
	Scenario 3	0.487	0.457	0.457	0.644	0.456	0.650
G322	Scenario 1	0.394	0.412	0.412	0.526	0.412	0.528
	Scenario 2	0.351	0.366	0.366	0.316	0.366	0.316
	Scenario 3	-	-	-	-	-	-
A267	Scenario 1	0.799	0.806	0.808	0.821	0.806	0.818
	Scenario 2	0.734	0.724	0.720	0.751	0.722	0.752
	Scenario 3	-	-	-	-	-	-
B601	Scenario 1	0.837	0.840	0.850	0.855	0.848	0.854
	Scenario 2	0.836	0.838	0.844	0.860	0.843	0.860
	Scenario 3	0.716	0.722	0.714	0.740	0.717	0.744
J567/01	Scenario 1	0.932	0.932	0.932	0.931	0.932	0.931
	Scenario 2	0.924	0.921	0.921	0.920	0.921	0.921
	Scenario 3	0.886	0.886	0.886	0.894	0.886	0.893

Table 10: Percentage of candidates whose estimated overall grade was the same as the actual grade

Unit/ Component	*Missing data generation*	*% achieving the same grade (Number of candidates = 60)*					
		Percentile	*z-score*	*Regression imputation*			
				OLS		*Quantile regression*	
				(a)	*(b)*	*(a)*	*(b)*
F212	Scenario 1	61.7	61.7	65.0	65.5	70.0	62.1
	Scenario 2	50.9	49.1	50.9	52.9	49.1	51.0
	Scenario 3	63.3	65.0	61.7	66.0	65.0	66.0
F213	Scenario 1	83.3	81.7	83.3	81.5	85.0	81.5
	Scenario 2	86.7	83.3	83.3	88.5	85.0	86.5
	Scenario 3	73.3	71.7	80.0	79.1	78.3	79.1
G322	Scenario 1	47.5	44.1	52.5	50.0	50.8	50.0
	Scenario 2	50.0	50.0	61.7	60.3	63.3	62.1
	Scenario 3	-	-	-	-	-	-
A267	Scenario 1	70.0	70.0	70.0	69.5	71.7	69.5
	Scenario 2	65.0	71.7	70.0	69.5	70.0	69.5
	Scenario 3	-	-	-	-	-	-
B601	Scenario 1	68.3	70.0	78.3	76.7	76.7	71.7
	Scenario 2	79.7	83.1	84.7	84.2	84.7	80.7
	Scenario 3	68.1	74.5	74.5	72.3	74.5	72.3
J567/01	Scenario 1	76.3	81.4	79.7	78.2	79.7	78.2
	Scenario 2	76.7	78.3	76.7	80.4	76.7	80.4
	Scenario 3	81.7	80.0	81.7	81.7	81.7	81.7

The percentages of candidates whose estimated overall grades were the same as the actual grades were very similar when using both the percentile and the z-score method.

There were, however, a couple of cases when the regression imputation provided the worst results and either the z-score or the percentile method was the best method in terms of preserving the actual grades. However, given the small numbers of cases used in each analysis, the differences are unlikely to be statistically significant.

Overall grade distributions for the qualifications including the units in Table 10 were also computed and compared with the actual grade

distributions. Table 11 shows the average absolute differences of the cumulative percentages between the actual and estimated grades.

Across half of the units and scenarios the regression imputation method (only results for models estimated using OLS regression and including additional data on students and schools are presented in the table) provided the (slightly) highest average, despite being the method providing the best results in terms of the percentage of candidates whose estimated overall grade was the same as the actual grade. However, because only a very small number of candidates have partial absences, the differences in the grade distributions should not be significant in practice.

The absolute differences varied slightly by grade. However this did not appear to be associated with a particular estimation method. For example, the absolute differences were not always bigger at grade A when marks were estimated by regression imputation. This is at odds with the work by Cheung (2009), who reported that the regression method had poor performance in estimating grades at both ends of the distribution.

Table 11: Average absolute differences of cumulative percentages between actual and estimated grades

Unit/ Component	*Missing data generation*	*Average absolute differences*		
		Percentile	*z-score*	*Regression imputation OLS (b)*
F212	Scenario 1	0.003	0.003	0.004
	Scenario 2	0.006	0.005	0.005
	Scenario 3	0.004	0.003	0.004
F213	Scenario 1	0.016	0.016	0.018
	Scenario 2	0.023	0.023	0.024
	Scenario 3	0.016	0.016	0.017
G322	Scenario 1	0.012	0.012	0.024
	Scenario 2	0.017	0.017	0.021
	Scenario 3	-	-	-
A267	Scenario 1	0.344	0.344	0.344
	Scenario 2	0.175	0.227	0.227
	Scenario 3	-	-	-
B601	Scenario 1	0.009	0.008	0.007
	Scenario 2	0.004	0.004	0.004
	Scenario 3	0.005	0.005	0.005
J567/01	Scenario 1	0.002	0.000	0.001
	Scenario 2	0.002	0.003	0.001
	Scenario 3	0.003	0.002	0.003

Conclusions

This work explored the numbers of unit/component marks in GCSE and AS/A level qualifications that were estimated by the OCR awarding body in the June 2015 session and compared three different methods (current and new) for estimating missing marks.

Very small numbers of marks were estimated. In particular, Ofqual found that 99.9% of A levels were graded by examiners in June 2014 (Espinoza, 2015). This was supported by the figures presented in this work, which highlighted that below 0.1% of the AS/A level scripts marked by OCR had estimated marks in June 2015.

Regarding the three methods compared in this research (z-score, percentile and regression imputation), regression imputation seemed to be the most accurate for estimating marks for all units/components and regardless of the different mechanisms used to create the partial absence data. When calculating grades based on estimated marks, the z-score and/or the percentile method were better in terms of preserving actual grades in a couple of instances. However, differences between methods were, in general, small.

In particular, the analyses presented here showed that:

- For the regression imputation method, the two different regression techniques considered (OLS and quantile) provided very similar results.
- For each of the regression techniques above, two different models were estimated. The first model only used the marks in the other units/components that counted towards the qualification; the second model included additional information about the candidates. The models with the additional data provided the best results.
- Although there was an effect of the estimation method on the accuracy of the estimated marks, the results of the analyses carried out here showed that the effect of the additional data was often bigger.
- When the marks in the units/components involved in the analysis do not correlate strongly (which is usually the case between coursework and written paper units/components), neither the z-score method nor the percentile method seems appropriate to estimate the missing marks. Regression imputation provides better estimates.
- Before estimating any missing marks, it seems worthwhile to identify which unit/component or combination of units/components should be considered and which ones should be discarded. Although this is a recommendation for all methods, it seems most relevant when using the z-score or the percentile methods.
- The percentages of candidates whose estimated overall grades were the same as the actual grades were very similar when using both the percentile and the z-score method. The regression imputation method provided, overall, the most accurate results.
- Work by Cheung (2009) found that the z-score method was better than regression (missing mark estimated based on marks in other units only) when comparing unit grades using an average of the absolute differences of cumulative percentages between actual and estimated grades. In a few instances, we found the same results for the overall qualification grade. However, because only a very small number of candidates have partial absences, the differences in the grade distributions would not be significant in practice.

Although regression imputation seems to have several advantages over the other two methods used currently by the UK awarding bodies offering GCSEs and AS/A levels, there is an important limitation to consider. If data is not available for a candidate in one or more of the variables included in the regression models (e.g., overall performance[5], average school performance), an estimated mark is not calculated. There are a couple of solutions in this instance. Firstly, the missing information can be estimated using statistical methods to handle missing data and, once available, the regression imputation proceeds as described in this article. An alternative is to use another method (e.g., z-score or percentile method) in those instances, or use in the imputation only those variables for which there is information.

5. Although, if a candidate has assessment scores from at least one other component then the overall performance measure should never be missing.

GCSEs and AS/A levels are currently being reformed, with many of the new reformed qualifications available for certification from June 2017. One of the main changes being introduced is the return to linear assessments. As a result, the JCQ has been recently working towards a common approach for how to calculate estimated marks in the new linear qualifications. Alternative methods such as the ones looked at in this research (e.g., z-score, percentile and regression imputation) have been considered in a variety of different research projects carried out by the different UK awarding bodies. The outcomes from such research did not show an outstanding method, but rather very small differences between them (this research shows just a marginal preference for regression imputation, with the performance of the z-score and percentile methods very similar). As the majority of the UK awarding bodies already use the z-score method for unitised specifications, it was agreed by the JCQ that it should be used for the new linear specifications from 2017 onwards.

References

Cheung, C.P. (2009). *Investigating different methodologies for calculating missing marks – examples using data from GCE AS new specifications (Economics and French)*. Internal Report. Cambridge: Oxford, Cambridge and RSA.

Espinoza, J. (2015). A-level results: exam boards 'guesstimating' students' grades. *The Telegraph*. (2015, August 01). Retrieved from: http://www.telegraph.co.uk/education/secondaryeducation/11777405/A-level-results-exam-boards-guesstimating-students-grades.html

JCQ (2016). *Access Arrangements and Reasonable Adjustments 2016–2017*. London: Joint Council for Qualifications. Available online at: https://examining.jcq.org.uk/exams-office/access-arrangements-and-special-consideration/regulations-and-guidance/access-arrangements-and-reasonable-adjustments-2016-2017

Linning, S. (2015). British exam boards forced to 'estimate' hundreds of A-level results each year as students' papers are 'lost in the post'. *Mail Online*. (2015, August 01). Retrieved from: http://www.dailymail.co.uk/news/article-3182362/British-exam-boards-forced-estimate-hundreds-level-results-year-students-papers-lost-post.html

On the reliability of applying educational taxonomies

Victoria Coleman Research Division

Introduction

Educational taxonomies are classification schemes which organise thinking skills according to their level of complexity. They provide a unifying framework alongside common terminology that can be used by educationalists. Primarily, most educational taxonomies focus upon thinking skills that fall within the cognitive domain, although some have also included other domains. Educational taxonomies can have a variety of different applications (Marzano, 2001). First, they can be used to analyse existing educational materials such as learning objectives, curriculum plans, lessons and assessments to ascertain which levels of thinking skills they encompass. Secondly, it is possible to use them as a framework when designing educational materials to ensure that the desired cognitive levels are targeted. They can also be adapted to form an assessment tool themselves, for example, by asking markers whether students have exhibited the required level of thinking during assessment activities. Finally, they can be used to ascertain whether corresponding curriculum objectives and assessment materials align, or whether there is a mismatch in the thinking levels that they are targeting. This can be done both in the context of designing new educational materials and in analysing pre-existing ones. Educational taxonomies can be applied in this way to a broad variety of educational contexts, being adapted according to the specific topic under investigation. This literature review will outline research investigating educational taxonomies and their use in terms of reliability.

Reliability

When discussing educational taxonomies and their application, it is important to consider reliability, as the value of different educational taxonomies is somewhat impacted by reliability constraints. There are various different types of reliability. In the subsequent literature review both inter- and intra-rater reliability are discussed[1]. The amount of consideration given to the rater reliability of educational taxonomies is highly variable and studies specifically investigating it in this context are sparse. There are three broad categories of techniques for assessing rater reliability: consensus estimates, consistency estimates, and measurement estimates (Stemler & Tsai, 2008). Within these there are a number of statistical methods that can be used in order to calculate reliability, and the technique which is selected depends on a number of factors such as the type of reliability being assessed and the nature of the data (McHugh, 2012; Stemler & Tsai, 2008). For example, a greater level of inherent dissimilarity between the items being categorised within a taxonomy will tend to lead to higher values of correlation-based reliability coefficients. However, if nearly all items being assessed fall into one or two categories, then there is little distinction between items and so less chance for raters to display a high correlation between their judgements. Conversely, if nearly all items are within a single category, simple measures (e.g., the percentage of times raters agree with one another) will appear high, as even random placement will result in a high level of agreement.

Within educational taxonomy research there has been a great deal of variation in the statistical measures used and in how the resulting reliability statistics have been interpreted. It must be noted that many of the research articles reviewed did not provide full details of the method used to calculate reliability, which limits our interpretation to some extent. Table 1 summarises the methods used by the studies in this review to calculate reliability.

1. Rater reliability is frequently referred to using different terms in the literature including coder, assessor and judge in place of rater, and terms such as consistency and agreement in place of reliability. However, for the purposes of this review, the term inter-rater reliability will be used.

Table 1: Reliability measures used by studies discussed in this article

Method	*Definition*	*Type of Reliability*	*Interpretation*
Percentage Agreement (PA)	Percentage agreement is a measure which is calculated as the percentage of times that two raters (or possibly groups of raters) gave the same rating.	Inter- and intra-rater reliability	≥70% acceptable reliability
Percentage Universal Agreement (PUA)	A variation on PA that measures the percentage of times that all raters gave the same rating.	Inter-rater reliability	≥70% acceptable reliability
Percentage Majority Agreement (PMA)	A variation on PA that measures the percentage of times for which the majority of raters gave the same rating.	Inter-rater reliability	≥70% acceptable reliability
Percentage of Partial Agreement (PPA)	Very similar to PMA. This is the percentage of instances in which there was partial agreement, defined as instances where 50% or more of the raters agreed. In this case exactly half of raters agreeing is counted as a positive outcome whereas for PMA it is not.	Inter-rater reliability	≥70% acceptable reliability
Kappa	Kappa coefficients include both Cohen's and Fleiss' kappa. Cohen's kappa can be used to assess the degree of consensus between two raters and whether it is above the level of agreement that would be expected to arise by chance alone. Fleiss' kappa is a similar measure which can be used when there are more than two raters.	Inter- and intra-rater reliability	Cohen's kappa (Cohen, 1960) < 0 poor agreement; 0.01–0.20 slight; 0.21–0.40 fair; 0.41–0.60 moderate; 0.61–0.80 substantial; 0.81–1.00 almost perfect (Landis & Koch, 1977). Fleiss's kappa <0.40 poor; 0.40–0.75 fair to good; >0.75 as excellent. (Fleiss, 1981).
Krippendorff"s alpha	Krippendorff's alpha can be used when calculating reliability for multiple raters with multiple possible ratings.	Inter-rater reliability	$\alpha \geq .800$ is good. $\alpha \geq .667$ is the lower limit for acceptable agreement (Krippendorff, 2004).
Correlations	Standard correlation coefficients such as Pearson's r and Spearman's Rho measure the association between two independent raters (or in some instances two groups of raters). In essence, they do not require that raters agree precisely on ratings, only that they place items in a similar rank order.	Inter- and intra-rater reliability	Values greater than 0.70 are typically considered acceptable levels of inter-rater reliability (Stemler & Tsai, 2008).
Cronbach's alpha	Estimates the expected correlation between the sum of scores across all raters and the (hypothetical) sum of scores across another group of raters of the same size.	Inter-rater reliability	Values greater than 0.70 are typically considered acceptable levels of inter-rater reliability (Stemler & Tsai, 2008).
Intraclass correlation coefficients (ICC)	ICCs attempt to overcome the limitations of other consistency estimates by taking into account both consistency and agreement of ratings. Essentially they measure the percentage of the variance across all ratings that is attributable to which item is being assessed (rather than which rater is doing the assessing).	Inter-rater reliability	Values greater than 0.70 are typically considered acceptable levels of inter-rater reliability (Stemler & Tsai, 2008).

Summary of the taxonomies mentioned in this review

This article presents an overview of reliability findings reported across a number of studies applying educational taxonomies. The concept of educational taxonomies was first introduced in 1956 with Bloom's Taxonomy of Educational Objectives, which provides a comprehensive system for classifying levels of thinking. It includes six categories of cognition, which are presented in a hierarchy of increasing complexity: knowledge, comprehension, application, analysis, synthesis and evaluation. Each category includes several subcategories. Bloom's taxonomy is also accompanied with examples of test items that belong to the different categories. Since its original introduction, it has been adapted and refined most notably by Anderson and Krathwohl (2001), whose revised taxonomy is widely used. Whilst the six cognitive categories remain in the revised version, some have been relabelled and it has moved away from the idea of a cumulative hierarchy and instead evolved into a two-dimensional framework with four knowledge categories added. Whilst a large number of alternative taxonomies have been developed, the vast majority of studies discussed in this review utilised Bloom's taxonomy, or adaptations of it (see Moseley et al., 2004 for a summary and review of educational taxonomies). Besides Bloom, there are numerous other educational taxonomies – those where research studies considering their reliability were found are also discussed and so are briefly outlined.

Another taxonomy that has been developed is the Structure of Observed Learning Outcomes (SOLO) taxonomy by Biggs and Collis (1982). Based on the stages outlined in Piaget's theory of cognitive development (1950), it includes five categories of understanding in a hierarchy of increasing complexity: prestructural, unistructural, multistructural, relational and extended abstract. An adaption of the SOLO taxonomy has divided the multistructural and relational categories into three subcategories each, resulting in nine SOLO levels in total (Burnett, 1999; Chan, Tsui, Chan, & Hong, 2002). A reflective thinking instrument developed by Kember (1999) can be used as an educational taxonomy to assess students' reflection and critical thinking skills. It is divided into two categories: non-reflective and reflective thinking. Non-reflective thinking is divided into habitual action and thoughtful action whilst reflective thinking is divided into reflection and critical reflection.

One of the studies in this review used Porter's taxonomy (Porter & Smithson, 2001a, 2001b). This was designed to enable standards and assessments in Mathematics and Science to be assessed. It includes three dimensions: topics, expectations of students' performance, and the modes of presentation. Each contains several subcategories. For example, the dimension of topics is a list of content areas within Mathematics and Science, with no hierarchical progression from one to the next. The final taxonomy considered is that of Marzano and Kendall (2006). This taxonomy comprises two dimensions: knowledge and mental processing. There are three knowledge domains: information, mental procedures and psychomotor procedures, with no hierarchy between these domains. Within the dimension of mental processing there are three hierarchical systems grouped into six levels. At the top is the self system, followed by the metacognitive system, with both comprising one level in the hierarchy. Following this is the cognitive system which is made up of four hierarchical levels: knowledge utilisation, analysis, comprehension and retrieval.

Literature examining the reliability of educational taxonomies

We found twenty-one studies[2] which considered reliability in the use of educational taxonomies – they are summarised in Table 2. These studies have utilised various educational taxonomies in a range of contexts and subject areas as well as employing several different measures of reliability.

The majority of studies in this review found evidence of moderate to high reliability when using educational taxonomies. The main exceptions to this were the research by Näsström (2009) and Karpen and Welch (2016). Näsström (2009) highlighted potential problems in both inter-

2. The articles were found through a Google Scholar literature search, with a list of established taxonomies searched alongside terms such as 'reliability', 'rater reliability' and 'rater consistency'. The reference lists of the initial papers that were found were then searched in order to find further relevant studies.

Table 2: Summary table of research studies examining the reliability of educational taxonomies

Study	*What was assessed*		*Which taxonomy*	*Raters*	*Inter-rater reliability (method in parenthesis)*	*Intra-rater reliability (time gap is in **bold**)*
Chan et al. (2002)	Term paper reports from 17 students		Modified nine category SOLO taxonomy Bloom's taxonomy Kember's Reflective Thinking Measurement Model (RTMM)	2 trained raters	Modified SOLO 0.60 (correlation between raters) Bloom r = 0.93 (correlation between raters) RTMM r = 0.87 (correlation between raters)	n/a
	Responses of 11 students to case study problems		As above – but with original 5 category SOLO taxonomy	2 trained raters	SOLO 0.66 (correlation between raters) Bloom's 0.68 (correlation between raters) RTMM 0.082 (correlation between raters)[a]	n/a
Crowe, Dirks, and Wenderoth (2008)	500 Life Science questions		Blooming Biology Tool–rubric based on adaptation of Bloom's taxonomy	3 raters	At least 2/3 agreed on 91% of the questions	n/a
	51 Life Science questions			36 students	98%(PMA) Additionally >80% agreed on 31/51 questions	n/a
Ebadi and Shahbazian (2015)	49 Iranian high school final exam questions		Bloom's taxonomy	2 panels of 2 researchers	0.87 (Cronbach's alpha)	**1 month later** 0.94 (Cronbach's alpha[b] for the first panel)
Edwards (2010)	Physics and Chemistry curriculum content and corresponding assessment papers		Revised Bloom's taxonomy	2 raters	0.97 for Physics curriculum objectives 0.98 for Chemistry curriculum objectives 0.88 for Physics assessment papers 0.92 for Chemistry assessment papers (method not specified)	n/a
Ewing, Foster, and Whittington (2011)	Classroom session in agricultural college	Professor discourse	Florida Taxonomy of Cognitive Behaviour – adaptation from Bloom's taxonomy	First rater: researcher Second rater: expert in cognition research	0.94 (method not specified)	**9 weeks later** 0.91 (method not specified
	Videotapes used for second rater and intra-rater reliability	Each professor question that elicited student engagement	Bloom's taxonomy	First rater: researcher Second rater not specified	0.93 (method unclear in article)	**3 weeks later** 0.84 (method not specified)
		Questions asked by students	Bloom's taxonomy	Second rater not specified	0.90 (method unclear in article)	**3 weeks later** 0.88 (method not specified)

a. The researchers stated 'the inter-rater reliability for Study 2 (the one which applied the modified version of SOLO with sub-levels) was higher than that of Study 1' (Chan et al., 2002, p.515). They suggested that this indicates that adding sub-levels increased inter-rater reliability. However, this seems to either be a misinterpretation or a reporting error as the modified SOLO was actually stated as being used in Study 1 and the unmodified SOLO showed the higher inter-rater reliability.

b. This does not particularly make sense as a measure of intra-rater reliability (in effect estimates the correlation of the sum of both measures with hypothetical set of two separate measures by the same individual). However, it is what was recorded by the author.

Table 2: Summary table of research studies examining the reliability of educational taxonomies *(continued)*

Study	*What was assessed*	*Which taxonomy*	*Raters*	*Inter-rater reliability (method in parenthesis)*	*Intra-rater reliability (time gap is in **bold**)*
	Course objectives	Bloom's taxonomy	First rater: researcher Second rater: expert in writing course objectives and cognition	0.98 (method not specified)	**3 weeks later** 0.92 (method not specified)
FitzPatrick and Schulz (2015)	165 educational outcomes statements and 182 corresponding statements from 2 units of Science curriculums from 4 jurisdictions	Revised Bloom's taxonomy	2 raters (An additional rater who was not included in reliability analysis)	80.4% for the outcomes (PA) 81.4% for the assessments (PA)	n/a
Karpen and Welch (2016)	Six questions from a teacher resource website, each targeted at a specific level of Bloom's taxonomy	Bloom's taxonomy	21 Pharmacy faculty members	0.25 (Krippendorff's alpha)	n/a
Leung (2000)	Responses from 79 students on an open ended DT item	SOLO taxonomy	1 researcher and 1 DT teacher, pre-marking meeting was held	0.49 (correlation between raters)	**Unknown time gap** 0.71 (correlation between the researchers marking and remark)
Mizbani and Chalak (2017)	57 speaking and listening activities from Iranian EFL Textbook Prospect 3	Bloom's revised taxonomy	Second rater for inter-rater reliability	0.92 (PA) on 14 of the activities	**2 weeks later** 0.98(PA) on random selection of 30 activities
Näsström and Henriksson (2008)	102 Swedish Chemistry upper secondary standards 58 assessment questions for upper secondary Chemistry	Bloom's revised taxonomy Porter's taxonomy (excluding modes of presentation domain)	2 raters	*Standards* 0.45 for Bloom's taxonomy (kappa) 0.07 for Porter's taxonomy (kappa) *Assessments* 0.36 for Bloom's taxonomy (kappa) 0.30 for Porter's taxonomy (kappa)	n/a
Näsström (2009)	35 Mathematics objectives for upper secondary schools in Sweden	Bloom's revised taxonomy	Panel of 4 assessment experts	26% (PUA, first occasion) 14% (PUA, second occasion) 46% (PMA, both occasions) 0.47 (kappa, first occasion) 0.41 (kappa, second occasion)	**2 to 3 months later** 51% (Average PA per judge) 12% SD 0.43 (kappa)
			Panel of 4 teachers	3% (PUA, first occasion) 11% (PUA, second occasion) 29% (PMA, both occasions) 0.15 (kappa, first occasion) 0.24 (kappa, second occasion)	**2 to 3 months later** 25% (Average PA per judge) 7% SD 0.18 (kappa)
Palmer and Devitt (2007)	33 MEQ's and 50 MCQ's from clinical undergraduate programme	Bloom's taxonomy	2 raters who then discussed and agreed a final rating	0.7 and 0.8 (kappa between each rater and the final agreed level MEQs) 0.7 and 0.8 (kappa, between each rater and the final agreed level MCQs)	n/a
Parham, Chinn, and Stevenson (2009)	84 statements in 24 transcripts from students' verbalisation when solving a Computer Science problem	Bloom's taxonomy	3 raters	89% (method not specified)	n/a
Plack et al. (2007)	308 reflective writing journal entries of medical students. These were assessed in terms of the highest level of cognitive processing that was displayed.	Three-level modified version of Bloom's taxonomy	3 raters	0.52–0.58 (kappa between pairs) 0.79 (ICC)	n/a
Razmjoo and Kazempourfard (2012)	One unit from Interchange EFL textbooks	Bloom's taxonomy	4 PhD student raters	0.972 (correlation of average rating of the PhD students with the researcher's rating)	**3 weeks later** 0.979 (correlation of average rating across all judges)
Rezaee and Golshan (2016)	41 questions in nationwide English exams in Iran	Bloom's taxonomy	2 raters	0.87 (correlation between two raters)	**Unknown time gap** 0.96 (1 rater's correlation with previous ratings)

Table 2: Summary table of research studies examining the reliability of educational taxonomies *(continued)*

Study	*What was assessed*	*Which taxonomy*	*Raters*	*Inter-rater reliability (method in parenthesis)*	*Intra-rater reliability (time gap is in **bold**)*
Riazi and Mosalendejad (2010)	Curriculum from 4 Iranian high school EFL textbooks	Bloom's taxonomy	No information provided	0.91 (method not specified)	**Unknown time gap** 0.98 (method not specified)
Teodorescu, Bennhold, Feldman, and Medsker (2013)	80 assessment questions from Physics textbooks	Physics adaptation of Marzano and Kendall's taxonomy	2 panels of 3 raters; each including 1 graduate student and 2 professors	0.75–0.85 (kappa between pairs on first panel) 0.70–0.82 (kappa between pairs on second panel)	**10 months later** 0.70–0.92 (kappa, between individuals on second panel)
Valcke, De Wever, Zhu, and Deed (2009)	282 messages as part of a collaborative learning group discussion task in Mathematics	Bloom's taxonomy	2 raters	0.95 (kappa)	n/a
van Hoeij, Hararhuis, Wierstra, and van Beukelen (2004)	179 short essay questions from 2 modules of a Veterinary course	Bloom's taxonomy	5 subject matter experts on first module	16% (PUA) and 57% (PPA) 34–57% (PA between each pair) <0.4 (kappa, between each pair)	n/a
			4 subject matter experts on second module	44% (PUA) and 28% (PPA) 61–77% (PA between each pair) 0.28–0.60 (kappa, between each pair)	n/a
			3 non-subject matter experts	*Module 1* 42% (PUA) and 49% (PPA) 50–71% (between pairs) <0.4 (kappa, between pairs) *Module 2* 50% (PUA) and 48% (PPA) 66–67% (between pairs) 0.55 (kappa, between pairs)	n/a
			All of the panels	*Inter-Group Reliability (non-experts vs experts)* Calculated 'modal taxonomic level' of each item for each panel group *Module 1* 65% (PA) 0.43 (kappa) *Module 2* 73% (PA) 0.63 (kappa)	n/a
Zheng, Lawhorn, Lumley, and Freeman (2008)	585 Biology exam questions (from Advanced Placement, undergraduate course, the MCAT, Graduate Record Examination and first year medical courses)	Bloom's taxonomy	3 experts in Biology education	0.53 (kappa) 0.68 (ICC)	n/a

and intra-rater reliability when using Bloom's taxonomy to assess the cognitive level of educational objectives, with none of the reliability findings showing more than moderate agreement. The findings are strengthened by the use of several different reliability measures, which have found consistent results. In particular, they found that across all of the measures, the teachers demonstrated lower reliability than the experts, which suggests there are differences as a function of the composition of the rating panels. In terms of differences between the two groups, the lower inter- and intra-rater reliability for the teachers may be related to the fact that they utilised the categories in Bloom's revised taxonomy to a greater extent and multi-categorised (categorising a single educational objective into multiple cognitive levels) to a lesser extent compared to the experts. This study involved the panels having group discussions about Bloom's taxonomy with examples presented to them before commencing their ratings. This is interesting given that training and practice was highlighted by many of the researchers as a potential avenue to improve reliability in the application of educational taxonomies. That said, conclusions about the impact of training and practice cannot be drawn from this study given that it was not examined experimentally.

Karpen and Welch (2016) also found low reliability when asking a panel of 21 faculty members to classify 6 exam questions. The researchers highlighted how this has implications for the use of Bloom's taxonomy and suggested that training of staff could be used to improve inter-rater reliability. That low reliability was found, when these questions had been purposefully written as examples of questions at specific levels in Bloom's taxonomy, is particularly concerning and perhaps also highlights challenges in using taxonomies to write questions at specific cognitive levels. It should also be noted that this study used a much greater number of raters than the other studies in this review. This therefore potentially raises the question as to whether the number of raters used impacts upon inter-rater reliability. That said, only six questions were rated and so this limits how much can be inferred from these results more generally.

Where reliability was investigated using standardised metrics such as kappa, alpha, correlation or ICC, the results tended to indicate acceptable to high reliability. However, our ability to draw conclusions about the reliability of taxonomies more generally is greatly limited by the fact that many of the studies reviewed did not specify the method that had been used to calculate reliability. Nevertheless, the high values reported by the majority of these studies do indicate a good level of reliability regardless of what measures were used to calculate them. Although, with the exception of Näsström (2009), all of the studies using PA (or variations such as PMA) found evidence of moderate to high reliability, these measures are limited in that they do not indicate how much agreement we could expect to find by chance alone. Consequently, given that there was a great deal of variability in the reported reliability found by studies using this measure, these findings must be interpreted with some caution when contributing to our overall conclusions about the reliability of educational taxonomies.

Overall, the majority of the studies provide evidence of moderate to good reliability when using educational taxonomies. In terms of inter- and intra-rater reliability, all of them considered inter-rater reliability with nine also examining intra-rater reliability. All of the studies examining intra-rater reliability found high reliability, with the exception of Näsström (2009). Whilst there were more measures of inter-rater reliability, these findings were more variable across the different research studies.

The majority of studies in this review used Bloom's taxonomy or adaptations of it, with just four including other taxonomies (Chan et al., 2002; Leung, 2000; Näsström & Henriksson, 2008; Teodorescu et al., 2013). Whilst this is unsurprising given the influence of Bloom in the field of education, the extent to which findings about the reliability of Bloom's taxonomy can be generalised to taxonomies more broadly is unclear. Consequently, other taxonomies would benefit from research being conducted to establish their reliability.

Areas for improving reliability

It is also useful to consider the way in which inter- and intra-rater reliability can be improved. The aforementioned studies have highlighted factors that influence reliability and which therefore offer a potential avenue for improvement.

Training and practice

The impact of training and practice on both inter- and intra-rater reliability was considered in a number of the research studies although none specifically examined their impact on reliability. Many of the studies either included some form of training (Chan et al., 2002; Näsström, 2009), or highlighted it as a potentially useful strategy for boosting reliability in future research (Karpen & Welch, 2016; Plack et al., 2007; van Hoeij et al., 2004). Training can be provided so as to ensure that raters are familiar with the taxonomy; it can also involve raters being given the opportunity to practise applying the taxonomy to sample materials, and having a group discussion to come to a consensus about how to interpret and apply the levels. Reliability can also be improved through providing examples of learning objectives of assessments that would fit into each taxonomic level. Some studies have demonstrated how a rubric with specific examples relevant to the topic area can be provided and used as a tool to support the application of educational taxonomies to both assessing and designing educational materials (Crowe et al., 2008; Lee, 2010). Whilst for some taxonomies, such as Bloom's, verb lists have been created to guide practitioners when applying them; evidence suggests that there is a great deal of variation in which verbs are aligned to specific levels and that individuals may interpret the meaning of different verbs at different levels (Stanny, 2016). Therefore it may be beneficial to include a group discussion as part of training, so that raters are able to develop a consensus in their interpretation and application of educational taxonomies. Thus, training and practice are potential ways in which reliability can be increased. As part of this, it is important to ensure that educational taxonomies and examples are clearly worded so as to reduce ambiguity and enhance reliability.

Rater variables

Characteristics of the rater also emerged as another factor which can impact upon rates of reliability, with differences found between experts and non-expert raters (e.g., Näsström, 2009). The number of raters used may influence inter-rater reliability, and there may be an optimum number of raters for achieving sufficient reliability whilst being practical in terms of constraints such as costs, time and finding sufficient number of raters, particularly where expertise is required. The number of raters also interacts significantly with characteristics of the raters, as the characteristics of the additional raters will impact upon the homogeneity of the group, with a homogenous group perhaps likely to show greater inter-rater reliability.

Taxonomy variables

The number of categories within an educational taxonomy was suggested as a potential factor impacting upon reliability by Chan et al. (2002), who suggested that adding sublevels to the SOLO taxonomy could increase inter-rater reliability by reducing ambiguity. Whilst it appears that their conclusion that subcategories increased reliability may be incorrect and based on a misinterpretation of the data (see footnote a), it would be useful for further research to investigate this and see if the number of categories and subcategories used impacts upon reliability. Although, of course, any conclusions on this matter would be hugely dependent upon the way in which reliability is defined.

Conclusion

Whilst very few studies were found which specifically examined the reliability of educational taxonomies, many studies did examine reliability to some extent. Although it is not possible to directly compare and summarise the findings across the studies, due to the different measures for assessing reliability, the majority of the studies discussed have provided evidence of at least moderate reliability, with evidence of poor reliability found only in a small number of instances – although of course studies showing poor reliability are less likely to get published. Inter-rater reliability has been looked at to a greater extent than intra-rater reliability. In the few studies that did consider intra-rater reliability, all but one found evidence of high intra-rater reliability. That said, many of these studies provided insufficient information about how reliability was calculated, such as failing to include information regarding which measure of reliability was used, and the time that elapsed between coding sessions. Consequently, the meaning and quality of the findings produced is sometimes unclear. Furthermore, this inconsistency in measurement places limitations on how far it is possible to compare reliability findings of different studies. Finally, whilst it seems that high reliability can be achieved using Bloom's taxonomy, and it can be hypothesised that high reliability can also be achieved when using other taxonomies, especially if

appropriate training and materials are used, there is insufficient research evidence to support or refute this hypothesis. In order to prove that research using other educational taxonomies can provide a sound evidence base for qualifications evaluation, comparability and development, further targeted studies will be necessary.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational objectives*. New York: Longman.

Biggs, J., & Collis, K. F. (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of the Observed Learning Outcome)*. New York.

Burnett, P. C. (1999). Assessing the structure of learning outcomes from counselling using the SOLO taxonomy: an exploratory study. *British Journal of Guidance & Counselling*, *27*(4), 567–580. Available online at: doi:10.1080/03069889908256291

Chan, C. C., Tsui, M. S., Chan, M. Y. C., & Hong, J. H. (2002). Applying the Structure of the Observed Learning Outcomes (SOLO) Taxonomy on Student's Learning Outcomes: An empirical study. *Assessment & Evaluation in Higher Education*, *27*(6), 511–527. Available online at: doi:10.1080/0260293022000020282

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297–334.

Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in Bloom: Implementing Bloom's Taxonomy to Enhance Student Learning in Biology. *CBE-Life Sciences Education*, *7*(4), 368–381. Available online at: doi:10.1187/cbe.08-05-0024

Ebadi, S., & Shahbazian, F. (2015). Exploring the Cognitive Level of Final Exams in Iranian High Schools: Focusing on Bloom's Taxonomy. *Journal of Applied Linguistics and Language Research*, *2*(4), 1–11. Available online at: http://jallr.com/index.php/JALLR/article/view/58

Edwards, N. (2010). An analysis of the alignment of the Grade 12 Physical Sciences examination and the core curriculum in South Africa. *South African Journal of Education*, *30*(4), 571. Available online at: http://www.scielo.org.za/scielo.php?pid=S0256-01002010000400005&script=sci_arttext&tlng=en

Ewing, J. C., Foster, D. D., & Whittington, M. S. (2011). Explaining Student Cognition during Class Sessions in the Context Piaget's Theory of Cognitive Development. *NACTA Journal*, *55*(1), 68–75. Available online at: http://www.jstor.org/stable/pdf/nactajournal.55.1.68.pdf?seq=1#page_scan_tab_contents

FitzPatrick, B., & Schulz, H. (2015). Do Curriculum Outcomes and Assessment Activities in Science Encourage Higher Order Thinking? *Canadian Journal of Science, Mathematics and Technology Education*, *15*(2), 136–154. Available online at: doi:10.1080/14926156.2015.1014074

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

Karpen, S. C., & Welch, A. C. (2016). Assessing the inter-rater reliability and accuracy of pharmacy faculty's Bloom's taxonomy classifications. *Currents in Pharmacy Teaching and Learning*, *8*(6), 885–888. Available online at: doi:10.1016/j.cptl.2016.08.003

Kember, D. (1999). Determining the level of reflective thinking from students' written journals using a coding scheme based on the work of Mezirow. *International journal of lifelong education*, *18*(1), 18–30. Available online at: doi:10.1080/026013799293928

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Sage.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *International Biometric Society*, *33*(1), 159–174. Available online at: doi:10.2307/2529310

Lee, H. A. (2010). *Thinking Levels in Christian Publishers' Elementary Reading Textbook Questions*. (Doctor of Education), Columbia International University.

Leung, C. F. (2000). Assessment for Learning: Using Solo Taxonomy to Measure Design Performance of Design & Technology Students. *International Journal of Technology and Design Education*, *10*(2), 149–161. Available online at: doi:10.1023/a:1008937007674

Marzano, R. J. (2001). *Designing a New Taxonomy of Educational Objectives*. California, USA: Corwin Press, Inc.

Marzano, R. J., & Kendall, J. S. (2006). *The New Taxonomy of Educational Objectives*. Thousand Oaks, CA: Corwin Press.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282. Available online at: http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=132393

Mizbani, M., & Chalak, A. (2017). Analyzing Listening and Speaking Activities of Iranian EFL Textbook Prospect 3 Through Bloom's Revised Taxonomy. *Advances in Language and Literary Studies*, *8*(3). Available online at: https://journals.aiac.org.au/index.php/alls/article/view/3527

Moseley, D., Baumfield, V., Higgins, S., Lin, M., Miller, J., Newton, D., & Gregson, M. (2004). *Thinking Skill Frameworks for Post-16 Learners: An Evaluation. A Research Report for the Learning and Skills Research Centre*. Retrieved from Regent Arcade House, 19–25 Argyll Street, London: http://files.eric.ed.gov/fulltext/ED508442.pdf

Näsström, G. (2009). Interpretation of standards with Bloom's revised taxonomy: a comparison of teachers and assessment experts. *International Journal of Research & Method in Education*, *32*(1), 39–51. Available online at: doi:10.1080/17437270902749262

Näsström, G., & Henriksson, W. (2008). Alignment of standards and assessment: A theoretical and empirical study of methods for alignment. *Electronic Journal of Research in Educational Psychology*, *6*(3), 667–690. Available online at: http://repositorio.ual.es/bitstream/handle/10835/565/Art_16_216_eng.pdf?sequence=1

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Medical Education*, *11*(11), 324. Available online at: doi:10.1186/1472-6920-7-49

Parham, J., Chinn, D., & Stevenson, D. E. (2009). *Using a Bloom's Taxonomy to Code Verbal Protocols of Students Solving a Data Structure Problem*. Paper presented at the 47th Annual Southeast Regional Conference, New York, USA.

Piaget, J. (1950). *The Psychology of Intelligence*. London: Routledge & Kegan Paul.

Plack, M. M., Driscoll, M., Marquez, M., Cuppernull, L., Maring, J., & Greenberg, L. (2007). Assessing Reflective Writing on a Pediatric Clerkship by Using a Modified Bloom's Taxonomy. *Ambulatory Pediatrics*, *7*(4), 285–291. Available online at: doi:http://dx.doi.org/10.1016/j.ambp.2007.04.006

Porter, A. C., & Smithson, J. L. (2001a). Defining, Developing and Using Curriculum Indicators. *CPRE Research Reports*. Available online at: http://repository.upenn.edu/cpre_researchreports/69

Porter, A. C., & Smithson, J. L. (2001b). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S. H. Furhmans (Ed.), *From the Capitol to the classroom: Standards-based reforms in the States* (pp.60–80). Chicago: National Society for the Study of Education, University of Chicago press.

Razmjoo, S. A., & Kazempourfard, E. (2012). On the Representation of Bloom's Revised Taxonomy in Interchange Coursebooks. *Journal of Teaching Language Skills*, *31*(1), 171–204. Available online at: http://www.sid.ir/En/VEWSSID/J_pdf/13112012660407.pdf

Rezaee, M., & Golshan, M. (2016). Investigating the Cognitive Levels of English Final Exams Based on Bloom's Taxonomy. *International Journal of Educational Investigations*, *3*(4), 57–68. Available online at: http://www.ijeionline.com/attachments/article/53/IJEI.Vol.3.No.4.06.pdf

Riazi, M. A., & Mosalendejad, N. (2010). Evaluation of Learning Objectives in Iranian High-School and Pre-University English Textbooks Using Bloom's Taxonomy. *The Electronic Journal for English as a Second Language*, *13*(4). Available online at: http://www.tesl-ej.org/wordpress/issues/volume13/ej52/ej52a5

Stanny, C. (2016). Reevaluating Bloom's Taxonomy: What Measurable Verbs Can and Cannot Say about Student Learning. *Education Sciences*, *6*(4), 37. Available online at: doi:10.3390/educsci6040037

Stemler, S. E., & Tsai, J. (2008). 3 Best Practices in Interrater Reliability Three Common Approaches. In J. Osborne (Ed.), *Best practices in quantitative methods*. Thousand Oaks: SAGE Publications, Inc.

Teodorescu, R. E., Bennhold, C., Feldman, G., & Medsker, L. (2013). New approach to analyzing physics problems: A Taxonomy of Introductory Physics Problems. *Physical Review Special Topics – Physics Education Research*, *9*(1). Available online at: doi:https://doi.org/10.1103/PhysRevSTPER.9.010103

Valcke, M., De Wever, B., Zhu, C., & Deed, C. (2009). Supporting active cognitive processing in collaborative groups: The potential of Bloom's taxonomy as a labeling tool. *The Internet and Higher Education*, *12*(3–4), 165–172. Available online at: doi:http://dx.doi.org/10.1016/j.iheduc.2009.08.003

van Hoeij, M. J. W., Hararhuis, J. C. M., Wierstra, R. F. A., & van Beukelen, P. (2004). Developing a Classification Tool Based on Bloom's Taxonomy to Assess the Cognitive Level of Short Essay Questions. *European Veterinary Education: Structuring Future Development*, *43*(3), 261–267. Available online at: doi:http://dx.doi.org/10.3138/jvme.31.3.261

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Application of Bloom's Taxonomy Debunks the "MCAT Myth". *Science*, *319*(5862), 414–415. Available online at:doi:10.1126/science.1147852

How much do I need to write to get top marks?

Tom Benton Research Division

Introduction

'How much am I supposed to write?' must be one of the most frequent questions students ask themselves when faced with an essay task. I remember this question being asked by someone in the class nearly every time such a task was set for homework at school, and my own children invariably ask me the same question every time I am encouraging them to do their homework. Despite the ubiquity of the question, clear answers are hard to come by. Teachers at my school would reply (rather unhelpfully) "how long is a piece of string?" whilst my response to my own children is rather more determined by how much I know they will be able to write before they start seriously complaining of fatigue than by any strong educational evidence.

There are good reasons not to answer this question. First and foremost is the fact that the quality of a response is not determined by the quantity of writing. For example, no published mark scheme for GCSEs will specify the amount that candidates are supposed to write but rather will rightly point markers towards the skill the assessment is supposed to be measuring; for example, in the case of English Literature[1], the extent to which candidates have identified the key features of the text they are studying and are able to communicate effectively. With these points in mind it is understandable if teachers want to make sure the student's efforts are focussed on producing a high-quality answer to the question and not on meeting some arbitrary target in terms of how much to write.

However, whilst this article is in no way arguing against the overriding importance of high-quality content, it is reasonable for students to want some guide to how much is expected in terms of length. An older *BBC Bitesize guide to English Literature GCSE* suggested that for a 45-minute examination students might have a target of roughly 450 words[2] – whilst also providing some more specific advice around time management and practice in structuring an essay. This article will supplement this advice by showing the amount of writing produced on average by candidates awarded different grades.

The relationship between the length of responses and the marks awarded to them has long been established within the field of automatic essay scoring. To take one example, Murray and Orii (2012) describe their own attempts to build a statistical model to achieve accurate essay scoring as part of a machine-learning competition. As a baseline comparator to their own technique, they present the correlation between predictions from a model based on essay length alone (both word count and character count) and the marks awarded to students. Across 9 different essay tasks, these correlations were all strongly positive, ranging from 0.50 to 0.82. Indeed, the extent to which automatic essay scoring algorithms can rely upon essay length has been criticised in research literature. For example, Perelman (2014, p.104) stated that "Automated Essay Scoring engines grossly and consistently over-privilege essay length in computing student writing scores" showing that, for the essays in this same competition, estimated scores from seven commercial vendors of automatic essay scoring were far more strongly related to word counts than was the case when human marking was used. However, there is no existing research linking the length of handwritten responses in GCSE examinations to the grades achieved by students.

Other research within the UK has investigated the average speed at which students can write under typical exam conditions. Such research is important for the purpose of determining the physical speed of writing below which a student may require further support by means of special considerations such as extra time or the facility to submit a typed (rather than a handwritten) essay as part of their examination. A review of this research is provided in Waine (2001). She reviewed 2 small-scale studies showing that in a free-writing task, where students had to decide what to write rather than simply copy it, students wrote on average between 14 and 18 words per minute. She also conducted her own study where, under examination conditions, 152 Year 10 (age 15) students were asked to write on the subject of 'My Life History' for a period of 30 minutes. Her results indicated that the mean writing speed of Year 10 students was 15 words per minute and that speeds between 10 and 20 words per minute were within the typical range. Similar research published by Patoss[3] (the professional association of teachers of students with specific learning difficulties) shows that, in a 20-minute free writing task, Year 10 students write at an average of 16 words per minute which rises to 17 words per minute for Year 11 students. Other research shows that when 16-year-old students are simply copying text they can write considerably even faster; at over 20 words per minute on average whilst writing neatly for 2 minutes, and at over 30 words per minute when writing as fast as possible (Barnett, Henderson, Scheib, & Schulz, 2009).

Overall, therefore, previous research has shown that the length of

1. See for example http://www.ocr.org.uk/Images/236719-mark-scheme-unit-a662-02-modern-drama-higher-tier-june.pdf (Retrieved 28 June 2017).
2. http://www.bbc.co.uk/schools/gcsebitesize/english_literature/prosejaneeyre/4prose_janeeyre_sprev1.shtml (Retrieved 28 June 2017).
3. https://www.patoss-dyslexia.org/SupportAdvice/InformationSheets/2012-09-02/Handwriting-Assessment/. (Retrieved 28 June 2017).

responses does have some association with achievement and also provided some norms around the possible writing speed of GCSE-taking-age children. However, none of these studies relate to performance in a real examination task. Thus, they do not provide any clue about how much writing is usually associated with achieving a high grade in a GCSE examination. The aim of this article is to fill this gap.

Basic method

As noted by Waine (2001), one of the main challenges with this type of research is the laborious task of manually counting the number of words written. To overcome this, building on work described in Benton (2017), the research for this article used computer processing of digital images of handwritten scripts to provide an estimate of how many words had been written. The basic process employed to count the number of words written on each page was as follows:

1. Use the dotted lines on the answer sheet to split the writing on the page into lines that can be processed separately (all essays included in this analysis were written on lined paper).
2. Remove any small objects (such as dots) from the image of each line. If, after this, there is no evidence of any ink remaining on the line then assume a word count of zero.
3. Within each line identify all clear horizontal gaps (i.e., horizontal spaces where there is no ink anywhere between the top and the bottom of the line being written on) and record the widths of these gaps.
4. Use cluster analysis to split these gaps between those that are likely to represent a break between words and those that are probably gaps between letters within the same word. In doing this it is assumed that any gaps wider than 5mm[4] must always represent a gap between words and that any gaps of less than 1mm must relate to a gap between letters within the same word.
5. The number of words on each line is now estimated as the number of between-word gaps on the line plus 1.
6. Add up the word counts across all lines on all pages within a candidate's examination booklet to produce a final estimated word count.

Further details on the processes involved in analysing images from examination scripts can be found in Benton (2017). The above approach was applied to a sample of 5,000 scripts from a 45-minute GCSE English Literature examination and the resulting word counts were linked with grades on the exam. However, before looking at the results of this analysis, it is first necessary to validate the word counting method itself.

Validation on a small scale example

In order to validate the word counting method above, the above process was applied to a sample of student responses to a short answer question from a GCSE Biology exam. The question itself asked "A supermarket is considering how they can make their shopping bags more sustainable. What is meant by sustainability?" and the answer space for students allowed them to write up to three lines of text in response. A random sample of 100 responses to this question was selected from amongst all

4. Actually 25 pixels within the resolution of images used for this analysis.

Word counts on question 'What is meant by sustainability?'

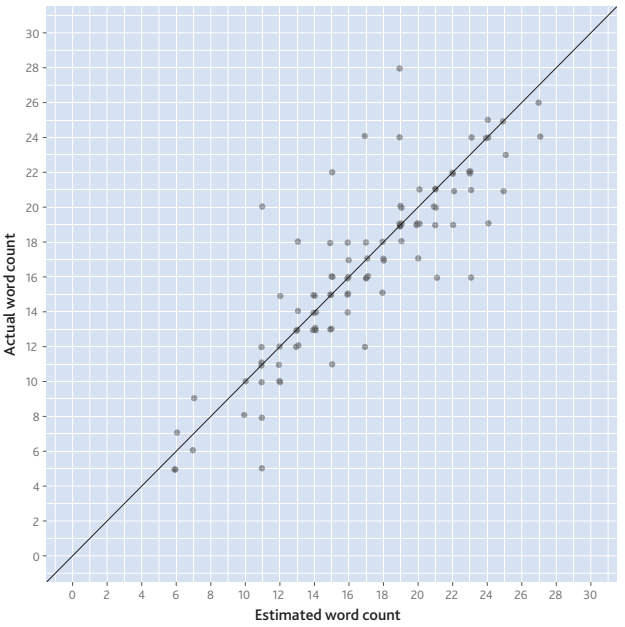

Figure 1: A comparison of estimated and actual word counts for the short Biology question

students taking the examination and a manual word count of each student's response was recorded. Then, the automatic process described above was applied to the same set of images and the estimated word counts were compared to the actual ones.

The results of the analysis are shown in Figure 1. As can be seen, the automated word counting mechanism was far from perfect but, nonetheless, did give a reasonable idea of the length of responses. The overall correlation between automated word counts and the actual length of responses was 0.87. Also, importantly, there was no evidence that the automated method was biased towards either over-counting or under-counting the true number of words. The actual mean number of words spent answering this question was 16.5 compared to a mean of 16.8 for the automatically estimated number. Similarly, the median of both the actual and the estimated word counts was 17. Further scrutiny of individual cases revealed that the automated word counts might be too low if candidates crossed out work and then rewrote sections of their answer over the top – thus obscuring any clear gaps between words. On the other hand, if a candidate's writing was too close to the line they were writing on (and perhaps dipped underneath this line) the algorithm may fail to include all of their writing within the image being analysed. This could lead to large horizontal gaps within words and, thus, the total number of words being over-estimated.

Notwithstanding these weaknesses, the analysis indicates that the automated method of generating word counts provides a reasonable basis for calculating how much candidates are writing in longer essays.

Word counts and grades for an English Literature examination essay

The analysis in this section examines GCSE English Literature essay responses from June 2014. In this particular examination, candidates were required to supply just a single essay response and were allowed a total of 45 minutes to complete their work. A random sample of 5,000 essays was selected for further analysis. The answer booklet was

restricted to six pages and a small minority of candidates where the number of archived scanned pages associated with their response differed from this was excluded from the analysis[5].

To begin with, for further validation, the automatic word counting process was applied to three pages from three different candidates taking this test. The average number of words per page from the automated process was found to match the actual average number of words showing that the process was generally suitable to be applied to full page responses.

Next, the automated word count process was applied to all essays. A total of 14 essays were removed from the analysis because the estimated word count was zero (this might be because the candidate's response was typed so was not within the standard answer booklet). A further two responses where the estimated word count exceeded 1,500 (which would imply the candidate wrote more than 30-words per minute throughout the entire exam) were also excluded from the analysis. The association between the estimated number of words written by each candidate and the grade they were awarded on this particular examination component is shown as a boxplot in Figure 2. The boxes in this plot indicate the inter-quartile range for the estimated word count within each grade with the central line denoting the median. The extra lines and dots show the full range of estimated word counts with the dots indicating outliers. Some summary statistics from this plot are provided in Table 1.

Table 1: Summary statistics for the relationship between estimated word counts and English Literature grades

Grade	*Number of candidates*	*Median estimated word count*	*Mean estimated word count*
A*	605	694	705.8
A	1008	637	652.0
B	1565	582	597.3
C	1009	517	538.8
D	493	492	500.7
E	142	450	460.4
U	162	370	383.5

Figure 2 shows a clear relationship between how much candidates wrote and the grades they were awarded. The correlation between estimated word counts and the marks awarded on the test (out of 49) was 0.46. The median number of words written by a grade A* candidate was 694 implying that they wrote around 15-words per minute in the exam, though, of course, they may not have used the entire time available in the exam for writing. Inspection of a few A*-graded essays of this length indicated that this relates to around five pages of writing. In contrast, the median number of words in a grade E essay was only 450 indicating 10 words were written per minute of the exam. In interpreting these numbers, it is important to remember that some of these candidates may have given up writing before the end of the available time.

Figure 3 shows the relationship the other way around, displaying the association between estimated word count and the number of marks awarded. The blue line shows how the mean number of marks awarded varied with the amount of writing. The dotted lines indicate the grade boundaries on the exam. Crucially, this shows that whilst candidates

5. A total of 27,351 candidate scripts for the exam were available within our digital script archive. The number of scanned pages in the archive matched the length of the answer booklet provided for 26,338 of these.

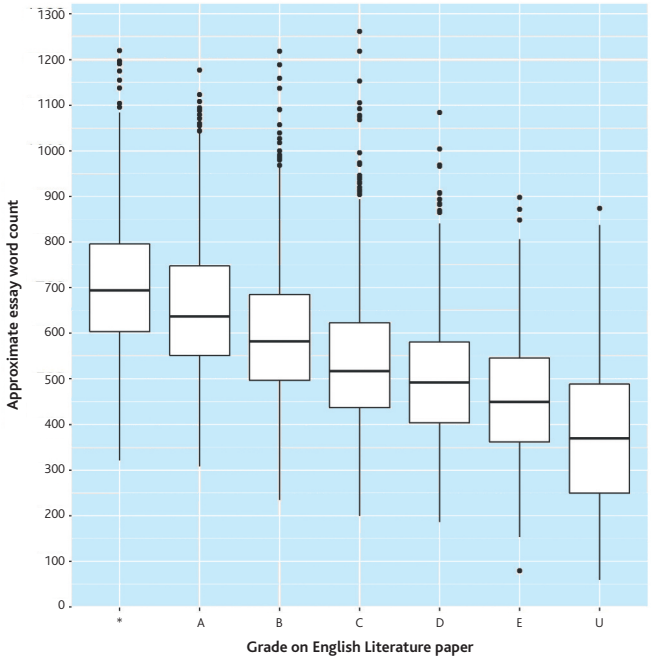

Figure 2: The relationship between word counts and achievement on the English Literature GCSE paper (The * represents the A* grade)

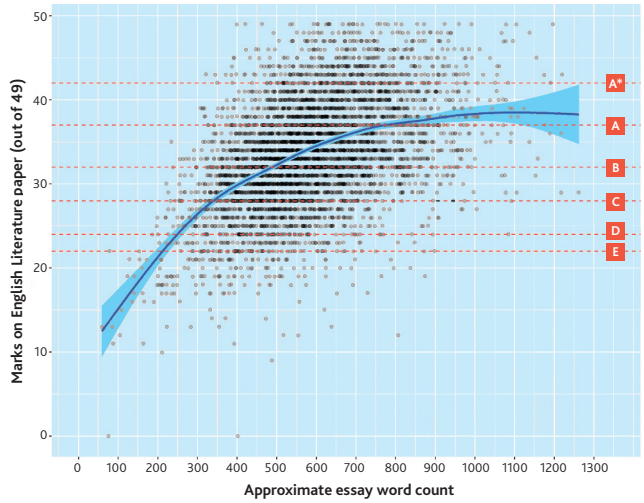

Figure 3: The relationship between estimated word count and mean mark

achieving the highest grades did tend to write more, a longer answer was by no means a guaranteed path to high marks. As can be seen, even for the longest essays, the average mark achieved by candidates never reached the top (grade A*) boundary. Indeed, the relationship between essay length and the mark awarded flattens off beyond 700 words indicating that there was no benefit in candidates writing extremely long responses.

At the other end of the spectrum the relationship is clearer. Nearly all responses of fewer than 200 words resulted in a grade U, suggesting that whilst very long answers are not necessary for a good mark, candidates must write enough to make sure that the examiner can recognise their knowledge at all. With this in mind it would be good advice for all candidates, even those who are not expecting to achieve the highest grades, to ensure that they produce at least a full page of writing in response to an English Literature exam question allowing 45 minutes to write an essay. It might also be noted that no candidate in the sample was awarded a grade better than a grade B without writing at least 300 words.

To complete the analysis, a brief manual inspection of some of the outliers in Figure 3 was conducted. Specifically:

- An inspection of the script awarded a grade E but where the estimated word count was below 100 revealed that a single candidate really did achieve a grade E with around half a page of writing.
- Inspecting the scripts for the two candidates apparently writing more than 1,200 words but only awarded a grade C revealed that both of them submitted 6 complete pages of writing. This reinforces the point that very long answers do not guarantee that a candidate will be awarded the highest marks.
- Inspecting the scripts for the 9 candidates awarded a grade A* but where the estimated word count was below 400 showed that in 8 out of 9 cases the candidates wrote less than 2.5 pages and in some cases less than 2 pages. For the other case (actually the grade A* candidate with the lowest estimated word count), the candidate had an unusually slanted writing style that probably obscured the gaps between words. Nonetheless, the other eight cases clearly show that it is possible to achieve the highest grades with fairly short answers.

Conclusion

This article has provided some fairly detailed information on the link between the amount candidates wrote for an English Literature essay and the marks they were awarded. As might be expected, there was a clear link, particularly at the lower end of achievement. This is no surprise as it is clearly impossible for candidates to be awarded the highest grades unless they provide enough material to demonstrate their skills to the examiner. With this in mind, if candidates are asked to spend 45 minutes answering an exam question they should aim to provide at least a page of writing in response and at least two pages (or thereabouts) if they want to have a chance of achieving any of the higher grades.

However, it is also very clear that the length of the response alone is insufficient to achieve a high mark. Beyond a certain essay length, the relationship between writing more words and achieving more marks flattened off. Thus, there is no evidence that writing extremely long answers makes a substantial difference to grade outcome, showing that quantity certainly does not trump quality. To reinforce this, we can note that inspection of individual essays revealed instances where, with well organised responses, students achieved all of the marks available on the exam with relatively short answers.

References

Barnett, A. L., Henderson, S. E., Scheib, B., & Schulz, J. (2009). Development and standardization of a new handwriting speed test: The Detailed Assessment of Speed of Handwriting. In *BJEP Monograph Series II, Number 6-Teaching and Learning*, 137–157. British Psychological Society. Available online at: https://doi.org/10.1348/000709909X421937

Benton. T. (2017). The clue in the dot of the "i": Experiments in quick methods for verifying identity via handwriting. *Research Matters: A Cambridge Assessment publication*, *23*, 10–16. Available online at: http://www.cambridgeassessment.org.uk/Images/375445-the-clue-in-the-dot-of-the-i-experiments-in-quick-methods-for-verifying-identity-via-handwriting.pdf

Murray, K.W., & Orii, N. (2012). *Automatic Essay Scoring*. Available online at: http://www.cs.cmu.edu/afs/cs.cmu.edu/Web/People/norii/pub/aes.pdf

Perelman, L. (2014). When the "state of the art" is counting words, *Assessing Writing*, *21*, 104–111. Available online at: https://doi.org/10.1016/j.asw.2014.05.001

Waine, L. (2001). Writing speed: What constitutes 'slow'? An investigation to determine the average writing speed of year 10 pupils. In Rose, R. & Grosvenor, I. (Eds.), *Doing research in special education: Ideas into practice*, 75–87. London: David Fulton Publishers.

Research News

David Beauchamp, Karen Barden and **Gill Elliott** Research Division, and **Gillian Cooke** Group Archives

50 years of research at Cambridge Assessment

This year, 2017, marks a particular milestone as the 50th anniversary of an established and permanent research unit which is now an integral part of Cambridge Assessment, previously the University of Cambridge Local Examinations Syndicate (UCLES).

On 1 August 1967 a meeting between representatives of three examination boards, (the University of Oxford Delegacy of Local Examinations, UCLES and the Oxford and Cambridge Schools Examination Board) agreed to the joint establishment of a research unit. A Cambridge location was chosen so that the unit could access the Cambridge Syndicate's IBM 360/30 computer - for one hour a day, possibly three, if evening work could be accommodated. The Test Development Research Unit (TDRU) was subsequently installed at 11 Station Road, Cambridge, on a 5-year lease.

After a great deal of fruitful research, TDRU was disbanded in 1985 as the tension between carrying out test development at speed and research at a sensible pace proved to be too much. However, a new research department was established within UCLES: the Council for Examination Development (CED). CED existed in the heady era of the development of the GCSE (introduced in 1988) and the IGCSE (introduced in 1989). Even the CED did not really go as far as the organisation wished in terms of pure research, so a study into research in assessment was commissioned. The results were the book *Assessment and Testing: A survey of research* by Robert Wood, and the establishment of the Research and Evaluation Division (RED) in 1994. RED was succeeded by Assessment, Research and Development (ARD) in the mid-2000s. At this time, we began publishing this journal *Research Matters*, a free biannual publication which allows us to share our research with the wider assessment community.

In 2015, the burgeoning possibilities of 'Big Data' led the division to establish a Data and Analytics team, tasked with operationalising analytics for the Group's exam boards and pioneering new applications of Data Science within Cambridge Assessment. The team introduced our series of *Data Bytes* to provide accessible visualisations of research findings to the wider public.

Just three years in, the TDRU Director reflected on his aim for 'No innovation without investigation'. Perhaps we will never know whether this was achieved, but it is a commendable ambition and, what is certain is that, it set a strong tradition for research within Cambridge Assessment which has continued ever since.

Conference presentations

Association for Language Learning, Nottingham, UK, March 2017

Carmen Vidal Rodeiro, Research Division: *The study of Modern Foreign Languages in England: uptake in secondary school and progression to Higher Education.*

Educational Collaborative for International Schools (ECIS) Leadership Conference, Barcelona, Spain, April 2017

Stuart Shaw, Cambridge International Examinations: *The assessment of collaboration: A 21st century response to a 21st century skill.*

National Conference on Student Achievement (NCSA), Texas, USA, June 2017

Stuart Shaw, Cambridge International Examinations: *Peer Review Submission from the stance and perspective of a UK-based international awarding body.*

British Education Studies Association (BESA), Liverpool, UK, June 2017

Jackie Greatorex, Research Division: *Two taxonomies are better than one: towards a method of analysing a variety of domains and types of thinking.*

European Conference on Social Media (ECSM) Conference, Vilnius, Lithuania, July 2017

Nicole Klir, Tom Sutch and James Keirstead, Research Division: *Tweeting about exams: social media discussion of British school exams.*

Journal of Vocational Education and Training (JVET), Oxford, UK, July 2017

Martin Johnson, Research Division and Tim Oates, Assessment Research and Development: *More like work or more like school? Insights into learning cultures from a study of skate park users.*

International Meeting of the Psychometric Society (IMPS), Zurich, Switzerland, July 2017

Tom Benton, Research Division: *Can artificial intelligence learn to equate?*

European Conference on Educational Research (ECER), Copenhagen, Denmark, August 2017

Nicky Rushton and Gill Elliott, Research Division: *Developing a framework for coding students' spelling errors in English.*

Carmen Vidal Rodeiro and Joanna Williamson, Research Division: *"Meaningful" destinations: using national data to compare progression to higher education, employment and training from different education pathways in England.*

Filio Constantinou, Lucy Chambers, Nadir Zanini and Nicole Klir, Research Division: *Formality in students' writing over time: empirical findings from the UK.*

Frances Wilson, OCR: *Reform of Practical Science Assessment in England: Impact on Teaching and Learning.*

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

Publications

The following articles have been published since *Research Matters*, Issue 23:

Bramley, T. (2017). Handbook of Test Development – review of Section 2. *Assessment in Education: Principles, Policy & Practice*. Advance online publication available at: http://dx.doi.org/10.1080/0969594X.2017.1297294

Bramley, T. (2017). Some implications of choice of tiering model in GCSE mathematics for inferences about what students know and can do. *Research in Mathematics Education*, *19*(2), 163–179. Available online at: http://dx.doi.org/10.1080/14794802.2017.1325775

Constantinou, F., Crisp, V., and Johnson, M. (2017). Multiple voices in tests: towards a macro theory of test writing. *Cambridge Journal of Education*. Advance online publication available at: http://dx.doi.org/10.1080/0305764X.2017.1337723

Darlington, E., and Bowyer, J. (2017). The Mathematics Needs of Prospective Geography Undergraduates. *Journal of Research in Social Sciences (JRSS)*, *5*(2), 11–32. Available online at: https://www.researchgate.net/publication/317381079_The_mathematics_needs_of_prospective_geography_undergraduates

Darlington, E. (2017). Coping styles of new undergraduate mathematicians. *Review of Science, Mathematics and ICT Education 11*(1), 5–17. Available online at: http://resmicte.lis.upatras.gr/index.php/review/article/view/2801

Darlington, E. and Bowyer, J. (2017). Students' views of A-level Mathematics as preparation for degree-level economics. *Citizenship, Social and Economics Education*, *16*(2) 100–116. Available online at: http://journals.sagepub.com/doi/abs/10.1177/2047173417716423

Darlington, E. and Bowyer, J. (2017). The role of 'extension papers' in preparation for undergraduate mathematics: students' views of the MAT, AEA and STEP. *Teaching Mathematics and its Applications*. Advance online publication available at: https://doi.org/10.1093/teamat/hrx009

Darlington, E. and Bowyer, J. (2017). Decision Mathematics as Preparation for Undergraduate Computer Science. *International Journal of Modern Education and Computer Science*, 9(4), 1–11. Available online at: DOI: 10.5815/ijmecs.2017.04.01

Gill, T., Vidal Rodeiro, C.L., and Zanini, N. (2017). Higher education choices of secondary school graduates with a Science, Technology, Engineering or Mathematics (STEM) background. *Journal of Further and Higher Education*. Advance online publication available at: http://dx.doi.org/10.1080/0309877X.2017.1332358

Johnson, M., Constantinou, F., and Crisp, V. (2017). How do question writers compose external examination questions? Question writing as a socio-cognitive process. *British Educational Research Journal 43*(4), 700-719. Available online at: http://dx.doi.org/10.1002/berj.3281

Shaw, S. D. (2017). Review of section 1 (foundations) – Handbook of Test Development. *Assessment in Education: Principles, Policy & Practice*. Advance online publication available at: DOI: 10.1080/0969594X.2017.1297293

Vidal Rodeiro, C.L. (2017). The study of foreign languages in England: uptake in secondary school and progression to higher education. *Language, Culture and Curriculum 30(3), 231–249*. Available online at: http://dx.doi.org/10.1080/07908318.2017.1306069

Vitello, S. and Williamson, J. (2017). Internal versus external assessment in vocational qualifications: a commentary on the government's

reforms in England. *London Review of Education*. Advance online publication available at: http://www.cambridgeassessment.org.uk/Images/internal-versus-external-assessment-in-vocational-qualifications.pdf

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

Reports of research carried out by the Research Division for Cambridge Assessment and our exam boards, or externally funded research carried out for third parties, including the regulators in the UK and many ministries overseas, are also available from our website: www.cambridgeassessment.org.uk/our-research/all-published-resources/research-reports/

Statistics Reports and Data Bytes

The ***Statistics Reports Series*** provides statistical summaries of various aspects of the English examination system, such as trends in pupil uptake and attainment, qualifications choice, subject combinations and subject provision at school. The reports, mainly produced using national-level examination data, are available in both PDF and Excel format on our website: www.cambridgeassessment.org.uk/our-research/all-published-resources/statistical-reports/

The most recent addition to the series is: *Statistics Report Series No. 113: Uptake of GCSE subjects in 2015, by alternative school type classifications.*

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

The following *Data Bytes* have been published since *Research Matters*, Issue 23 – Interactive graphics are marked with (I):

- April 2017: *Do grades in one GCSE subject predict results in another?* (I)
- June 2017: *The most popular non-compulsory GCSE subjects in the period 2005–2014*
- July 2017: *European participation in employer-sponsored vocational training*
- September 2017: *Progress towards universal primary education.*

All details are correct at the time of going to print in October 2017

Contents / *Issue 24* / Autumn 2017

2 **Undergraduate Mathematics students' views of their pre-university mathematical preparation** : Ellie Darlington and Jessica Bowyer

11 **Question selection and volatility in schools' Mathematics GCSE results** : Cara Crawford

17 **Utilising technology in the assessment of collaboration: A critique of PISA's collaborative problem-solving tasks** : Stuart Shaw and Simon Child

23 **Partial absences in GCSE and AS/A level examinations** : Carmen Vidal Rodeiro

30 **On the reliability of applying educational taxonomies** : Victoria Coleman

37 **How much do I need to write to get top marks?** : Tom Benton

42 **Research News :** David Beauchamp, Karen Barden, Gill Elliott and Gillian Cooke

Cambridge Assessment
1 Hills Road
Cambridge CB1 2EU
United Kingdom

+44(0)1223 552666
researchprogrammes@cambridgeassessment.org.uk
www.cambridgeassessment.org.uk

9274642234

ISSN: 1755–6031