References

Adams, R. J., Wu, M. L., and Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, *72*(4), 547–573.

Ahmed, A. and Pollitt, A. (2011) Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy and Practice*, *18*(3), 259–278.

Baird, J. A., Beguin, A., Black, P., Pollitt, A. and Stanley, G. (2011). The Reliability Programme: Final Report of the Technical Advisory Group. In: *Ofqual's Reliability Compendium* (Chapter 20). Coventry: The Office of Qualifications and Examinations Regulation.

Bramley, T. (2001) The Question Tariff Problem in GCSE Mathematics. *Evaluation and Research in Education*, *15*(2), 95–107.

Fortune, T. W. and Tedick, D. J. (Eds.) (2008) *Pathways to multilingualism: Evolving perspectives on immersion education*. Clevedon, England: Multilingual Matters, Ltd.

Fowles, D. (2009) 'How reliable is marking in GCSE English?' *English in Education*, *43*(1), 49–67.

Fransella, F. Bell, R. and Bannister, D. (2004) *A Manual for Repertory Grid Technique (2nd Edition)* Chichester, UK: John Wiley and Sons.

Galaczi, E. D., ffrench, A., Hubbard, C. and Green, A. (2011) Developing assessment scales for large-scale speaking tests: a multiple-method approach, *Assessment in Education: Principles, Policy and Practice*, *18*(3), 217–237.

Jankowicz, D. (2004) *The Easy Guide to Repertory Grids*. Chichester, UK: John Wiley and Sons.

Kelly, G. A. (1955/1991) *The Psychology of Personal Constructs*. London, UK: Routledge.

Landfield, A. W. (1971) *Personal Construct Systems in Psychotherapy*. Chicago, USA: Rand McNally.

Linacre, J. M. (2005). *A User's Guide to FACETS Rasch-Model Computer Programs*. Available at: www.winsteps.com

Lumley, T. (2002) Assessment criteria in a large-scale writing test: What do they really mean to raters? *Language Testing*, *19*(3), 246–76.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika* *20*(2), 149–174.

Milanovic, M. and Saville, N. (1996) Performance testing, cognition and assessment: selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem. *Studies in Language Testing 3*. Cambridge, UK: University of Cambridge Local Examinations Syndicate/Cambridge University Press.

Pollitt, A. (1991). Response to Alderson, Bands and scores. In J. C. Alderson & B. North (Eds.), *Language Testing in the 1990s*. London, UK: MacMillan.

Shaw, S. D. and Weir, C. J. (2007) Examining Second Language Writing: research and practice. *Studies in Language Testing 26*. Cambridge, UK: University of Cambridge Local Examinations Syndicate/Cambridge University Press.

Sweiry, Z., Crisp, V., Ahmed, A, and Pollitt, A. (2002) *Tales of the Expected: The Influence of Students' Expectations on Exam Validity*. British Educational Research Association Conference, Exeter, UK, September 2002.

Tisi J., Whitehouse, G., Maughan S. and Burdett, N. (2013) *A Review of Literature on Marking Reliability Research* (Report for Ofqual). Slough: National Foundation for Educational Research (NFER).

Upshur, J. and Turner, C. (1995) Constructing rating scales for second language tests. *ELT Journal*, *49*(1), 3–12.

Weir, C. J. (2003) A survey of the history of the Certificate of Proficiency in English (CPE) in the twentieth century. In C.J. Weir & M. Milanovic (Eds.). Continuity and Innovation: A History of the CPE Examination 1913–2002. *Studies in Language Testing 15*. Cambridge, UK: University of Cambridge Local Examinations Syndicate/Cambridge University Press.

Maintaining test standards by expert judgement of item difficulty

Tom Bramley Research Division and **Frances Wilson** OCR (The study was completed when the second author was based in the Research Division)

Introduction

This article describes two methods for using expert judgements about examination questions (items) to arrive at a cut-score (grade boundary) on a new examination paper where none of the items has been pre-tested. We wanted to see if we could exploit the wealth of data about item difficulty that has been available in the years since the majority of papers have been marked (scored) on-screen.

The General Certificate of Secondary Education (GCSE) and the General Certificate of Education GCE Advanced level (A level) are high-stakes curriculum-based examinations taken at age 16 and 18 respectively by pupils in England. They are offered by three Awarding Organisations (AOs), and schools can decide which AO's exams they enter their pupils for. Outcomes are reported on a grade scale (A* to G at GCSE; A* to E at A level, with U indicating 'ungraded' for both). From 2017, reformed GCSEs in England will be graded on a 1–9 scale. The full assessments normally consist of several components (e.g., written examination papers, practical or coursework assessment, portfolios, speaking tests, musical performances etc.). The assessments are usually graded at component level, and the overall grade is determined by aggregation rules which can vary considerably depending on the structure of the assessment (e.g., whether the assessment is 'linear', where all components are taken at the end of the course, or 'modular', where assessment units can be taken at various stages throughout the course). At component level, the grading process involves establishing the cut-scores (grade boundaries) on the raw mark scale that define the ranges of raw scores mapping to each grade.[1] A regulatory code of practice (Office of Qualifications and Examinations Regulation [Ofqual], 2011) sets out the mandatory aspects of this process, which requires the AOs to consider a variety of sources of evidence. Benton and Bramley (2015) show that these sources of evidence can be broadly classified as: i) evidence about the ability of the cohort of examinees; ii) evidence about the difficulty of the examination; and iii) evidence about the quality of work produced in the examination.

Setting the grade boundaries is essentially a standard-maintaining process (as opposed to a standard-setting process) where the aim is for

[1] Only particular 'key boundaries' are established by the 'Awarding Committee' – the other boundaries are derived from these by interpolation rules. At A level, the key boundaries are at grades A and E.

the resulting grades to have the same meaning over time and across AOs. However, the decision on where to locate the grade boundaries has to combine the evidence from the three sources – which is not an easy task since they can relate to different conceptions of what it means to maintain a standard. The conceptual and practical problems created by this are well documented (e.g., Baird, Cresswell & Newton, 2000; Newton, Baird, Goldstein, Patrick & Tymms, 2007; Coe, 2010).

Traditionally the first and third of these sources have been the most dominant, and in recent years the first source (in the form of the 'comparable outcomes' method [Benton & Lin, 2011; Taylor, 2014]) has particularly constrained the possible locations of the boundaries. The second source of evidence (about the difficulty of the examination) has played a more minor role. This is partly because the high-stakes nature of the assessment makes pre-testing and re-use of items[2] impractical for security reasons (ruling out statistical evidence about item difficulty), and partly because there seems to be some scepticism about the ability of experts to provide accurate and reliable information about difficulty based on their informed judgements about examination questions.

This scepticism is based on the well-known method-dependence in the results of various item-based standard-setting methods (e.g., Glass, 1978); the variability in results within a given standard-setting method attributable to the judges (e.g., Clauser, Margolis & Clauser, 2014); the possibility that the expert judgements have a different implied scale unit from the student responses (Humphry, Heldsinger & Andrich, 2013); and the fact that there is often poor absolute agreement between judged item difficulty and empirical item difficulty (e.g., Bejar, 1983; Impara & Plake, 1998). However, there is also evidence in the research literature that in some circumstances there can be reasonable agreement between judged and empirical difficulty, particularly when judgements of experts are pooled; when those making the judgements are properly trained; when there is empirical data for judges to 'anchor' their judgements; and when judgements of difficulty are relative, rather than absolute (e.g., Brandon, 2004; Hambleton & Jirka, 2006; Attali, Saldivia, Jackson, Schuppan & Wannamaker, 2014). Most of the research on judgement of difficulty has been in the context of standard-setting methods for tests comprising objective (usually multiple-choice) dichotomous items. It is therefore an open question as to whether, and how best, expert judgement of difficulty can be used in the standard-maintaining context of GCSEs and A levels for those components where the majority of items are polytomous, short answer questions.

In this standard-maintaining context, a large and as yet largely untapped source of relevant information is available to guide the experts in their judgements of difficulty – statistical information about the performance of examinees on each item of previous versions of the component. The study reported here involved using this information in two different ways to derive estimates of the grade boundaries.

The first way was closely related to the Angoff standard-setting method and its extension to polytomous items – the 'mean estimation' method (e.g., Loomis & Bourque, 2001). If experts are able to estimate the mean score of examinees on the borderline of a particular grade for each item, then summing these estimates across the items would give an estimate of the grade boundary for that grade. It is possible to provide the experts with the actual mean scores of boundary examinees for each item on previous versions of the relevant component. Therefore their estimates of mean scores for boundary examinees on the new version of the component can be guided by their judgements of similarity of new items to previous items. The advantage of this method is that it can be applied well before the examination is taken – that is, it does not require any statistical information from the examination itself. A potential disadvantage is that it still requires numerical estimates from the experts which, as we have shown, there are reasons to doubt that they can make reliably enough.

The second way was technically more complex, but required less from the experts. It was based on an idea first suggested in Bramley (2010):

> *On [this] approach … the awarding panel would identify questions on the current paper which are similar enough to questions on a previous paper for it to be reasonable to expect performance on them to be equivalent. Now the argument would be along the following lines: 'Last year the borderline grade C examinees (with a test score of 40) averaged 1.2 out of 2 on question 7a, which required them to label a diagram of a cell. This year's question 3b was practically identical, and examinees who averaged 1.2 out of 2 scored 42 on the test overall, suggesting a mark of 42 would be appropriate for this year's boundary.*
>
> (Bramley, 2010, p.35).

Thus here the task was merely to find item(s) on previous versions that were similar or identical to each item on the current version. Empirical item characteristic curves (EICCs) were created for each item on previous versions of the component, and for each item on the new component (as soon as the data was available). These EICCs were smoothed plots of item score against total score using the TRANSREG procedure in SAS® software with a smoothing parameter of 50.

Estimates for a particular grade boundary on the new version were derived from the following steps:

1. Find (from the relevant EICC plot or by tabular interpolation) the item score corresponding to the grade boundary on each previous item judged to be similar or identical to a new item;
2. Find (from the relevant EICC plot or by tabular interpolation) the total score on the new test corresponding to each item score identified in step 1;
3. Average the total scores obtained in step 2.

The process is illustrated graphically for a single item in Figure 1.

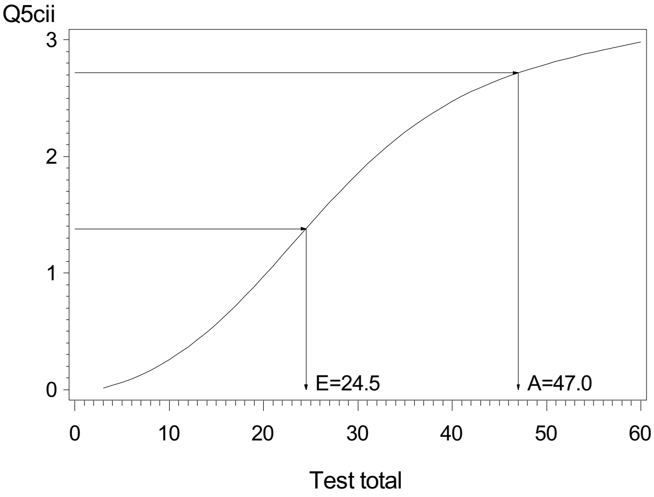

Figure 1: Smoothed plot of item score against test score for the new test

[2] In this article, an 'item' is a subpart of a larger question. For example, Q1a (part of Q1) would be considered to be an item. 'Question' and 'item' are used interchangeably depending on the context.

In Figure 1 the previously obtained (smoothed) mean score on the item judged to be similar/identical to Q5cii for grade A boundary examinees was 2.72 and for grade E examinees it was 1.38. Reading across from these values on the y-axis and down to the x-axis gives the A and E boundaries implied by performance on this 'pseudo-anchor' item. In Figure 1 these are 47.0 and 24.5 respectively. Averaging these implied boundaries at each grade across all similar/identical items identified by the judges gives the A and E boundaries produced by this method. Since this method requires statistical data from the new test, it can only be carried out after the examination has been taken and marked (scored).

The present study applied these methods to an A level Chemistry component in order to estimate grade boundary locations for the June 2014 paper. Interest focused on the variability of results across experts and the agreement of the resulting grade boundaries with the actual grade boundaries that were eventually set. The procedures and results are described below. The discussion relates the above methods to existing methods in the standard-setting and maintaining literature and considers further their strengths and weaknesses.

Method

The A level Chemistry component was chosen for the research because it had a large, stable entry, with reliable statistical data at item level available for more than six previous versions.[3]

Participants

Because of the need to maintain security of the examination materials, only two experts – the Principal Examiner (Expert 1 [Ex1]) and the Chief Examiner (Expert 2 [Ex2]) – were used. They had already seen the June 2014 paper (because they were involved in setting the questions) so the exercise could be completed before the date of the live examination. The second author (Au) also completed the task to allow comparison between expert and non-expert judgements; her highest qualification in Chemistry was A level, though she had recently worked on a number of research projects relating to Science qualifications.

Materials

The experts were sent the following materials:

- Past question papers and mark schemes (scoring rubrics) from each previous version from January 2011 to June 2013 (six papers in total);
- The question paper and mark scheme to be taken in June 2014;
- A spreadsheet which listed the (smoothed) mean mark achieved by examinees on the grade A and grade E boundary on each subquestion on each previous paper. The specification reference[4] was provided for each subquestion;
- A spreadsheet listing each subquestion and the specification reference for each subquestion on the Summer 2014 paper, for participants to fill in their responses.

Task

The experts were asked to estimate the mean marks that would be obtained on each subquestion of the June 2014 paper by examinees on the grade E boundary, and examinees on the grade A boundary.

They were instructed to use the specification reference information provided for the Summer 2014 paper to identify past questions which assessed similar or identical material to use as the basis for their judgement. If they could find a question that was identical, or nearly identical, in the past papers to the Summer 2014 question, they were instructed to use its previous empirical values of facility for A and E boundary examinees as their estimate. If questions which were similar, but not identical, could be found, then they were instructed to use those past values as a basis for their estimate, but to modify their estimate according to their judgement of the effect of any differences on difficulty. Where no similar questions could be found, they were asked to use their own judgement. They were asked to explain the rationale for their decisions.

Results

Of the 33 subquestions on the June 2014 paper, there was only 1 where neither expert was able to identify anything similar or identical in any of the previous 6 papers. For 2 of the subquestions, 10 and 11 similar previous subquestions were identified. Most commonly between two and six similar subquestions were identified. The judges differed considerably in how many similar questions they identified in total – Ex1 found 38, Ex2 found 90, and Au found 30.

Table 1: Agreement between the experts in number of instances of similar questions identified

Judge	*No. of questions*
Ex1 only	4
Ex2 only	44
Au only	3
Ex1 and Ex2*	29
Ex1 and Au*	16
Ex2 and Au*	30
Ex1, Ex2 and Au	14

*Regardless of whether the third judge identified that question.

Table 1 shows that, given that Ex2 identified many more similar questions than the other two judges, it was rare for the other two to find a question that he had not identified. There were 14 instances where all 3 judges agreed, representing 12 questions on the June 2014 paper. (For two questions, all three judges agreed there were two similar questions that had been asked before).

Mean estimation method

Figure 2 shows the agreement between the two experts' judgements at grade A and E. The estimates of mean marks have been scaled by the number of marks available for the subquestion (in other words the graphs show the estimated facility values) in order to highlight more clearly where there were differences between the experts. It is clear from Figure 2 that Ex1 generally estimated higher values than Ex2 at both A and E, although the judgements at both grades were reasonably well correlated (see Table 2).

[3] This component was available to examinees in January or in June. A completely new version was created each time.

[4] A code indicating the area of the specification (syllabus) relevant to the question. For example, 1.2.1j referred to the subsection of the specification about electron structure (1.2.1) and the 'assessable learning outcome' j which was 'classify the elements into s, p and d blocks'.

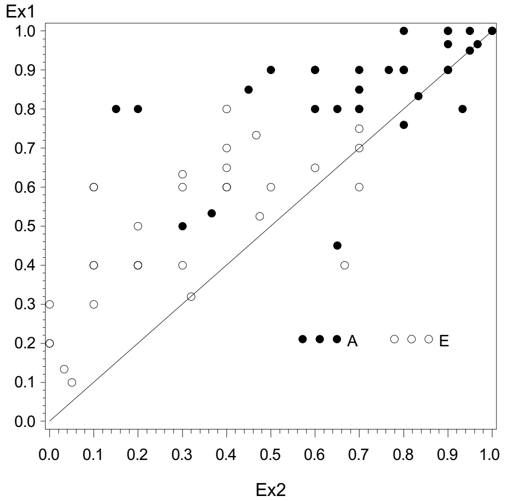

Figure 2: Estimated facility value for Expert 1 *v* Expert 2 at grade A (dots) and grade E (circles)

Table 2: Inter-correlations of estimated and actual facility values for boundary examinees. Grade A above and right of the main diagonal, grade E below and left. (N=33)

	Ex1	*Ex2*	*Au*	*Actual*
Ex1		0.60	0.20	0.41
Ex2	0.77		0.13	0.66
Au	0.51	0.56		0.14
Actual	0.72	0.76	0.52	

Table 2 shows that there was better agreement among the judges at grade E than grade A. It is particularly noticeable that there was a higher correlation between the two experts at both grades than between either of them and Au, although this represents expert agreement about relative difficulty rather than absolute difficulty, as seen in Figure 2. The correlation of the expert judgements with the actual values later obtained when the paper 'went live' was quite high for Ex2 at both grades, but only at grade E for Ex2, whose correlation with the actual values was only 0.41 at grade A.

Table 3: Grade boundaries implied by the judgements

Judge	*Grade A sum*	*Grade A*	*Grade E sum*	*Grade E*
Ex1	50.6	51	30.2	30
Ex2	44.2	44	21.8	22
Au	49.7	50	27.2	27
Mean (all)	48.2	48	26.4	26
Mean (experts only)	47.4	47	26.0	26
Actual boundary		46		26

At grade A the boundaries implied by the judgements of the two experts were 7 marks apart and at grade E they were 8 marks apart, a rather discouraging finding given an average grade bandwidth (difference between grade boundaries on the raw mark scale) of around 5 marks on previous versions of this component. However, the mean of their judgements did equal the eventual actual boundary at grade E and was only 1 mark too high at grade A. The boundaries implied by the researcher's (non-expert) judgements were between those of the experts, and did not significantly affect the mean at grade E, but raised it at grade A to a value 2 above the actual boundary.

Similar items method

Applying the second method for deriving grade boundaries involved deciding which items on the June 2014 test should be deemed similar enough to previous items to justify using the previous statistics. An initial list contained 15 items from the June 2014 paper where both experts and the researcher had identified the same similar previous question. The first criterion we used for selecting similar questions from this initial list was to choose questions with the same maximum mark as the previous question and where at least one of the expert judges had used the same value as the previous value as their estimate (i.e., at least one expert thought the difficulty would be the same). This criterion produced slightly different lists of questions for grade A and E (because the experts could have used the same value as the previous question for one grade but modified the value for the other). At grade A, 9 items worth 17 marks met the criterion, and at grade E 8 items worth 15 marks did.

Next, we tried a stricter criterion for anchor item selection, choosing only those items where all three judges had agreed and where all had used the same values as previously for both the A and E boundaries. This only identified two items worth 4 marks in total.

Finally we tried just using the judgements of Ex2 (who had identified the most similar items, and whose correlations with the actual values were highest at both grades), taking only those items where he had used the same value as the previous statistics (i.e., that he judged to be of identical difficulty to a previous question). This gave 11 items worth 20 marks in total. The text of these items, but not their layout, is shown in Table 5 in the Discussion section.

Table 4: Grade A and E boundaries implied using three different criteria for identifying similar items

	No. of similar items	*No. of marks*	*A*	*A (rounded)*	*E*	*E (rounded)*
Criterion 1	8 (A)/9 (E)	15/17	46.3	46	24.2	24
Criterion 2	2	4	48.7	49	27.0	27
Criterion 3	11	20	46.8	47	25.4	25
Actual				46		26

Table 4 shows that all three criteria for identifying similar previous items led to similar estimates for the grade boundaries, and that in all cases they were close to the actual boundary. This suggests that the second method of deriving boundaries may be a better way to use the data available.

Discussion

The aim of this study was to investigate two methods for deriving grade boundaries on an exam paper, using expert judgements about the questions. The first method could be characterised as statistically-informed expert judgement about question difficulty. It transplants the Angoff standard-setting method into a standard-maintaining context where the experts can use statistical information about performance of grade boundary examinees on items in past versions (forms) of the test to inform their judgements about the likely performance of grade boundary examinees on the new test. The main advantage of the method is that it provides a source of evidence about the difficulty of the new test which is independent of any data about the performance of examinees on the new test, and hence can be applied before the test is taken. Furthermore, the previous statistical information can be used intelligently by the experts to guide their judgements based on how similar they think each new item is to one that has been asked in the past. The results of this study suggested that the first method would need to involve more expert judges to reduce the impact of variability among the judges on the final outcome. Although the outcomes from this study were close to the actual boundaries, this may have been due to luck considering how far apart the two experts' judgements were (in absolute terms).

The second method could be characterised as non-parametric Item Response Theory (IRT) common item equating using expert judgement of item similarity to define pseudo-anchor items. As far as we are aware, this is a new method, although the idea of using smoothed EICCs has appeared in a recent article by Zu and Puhan (2014), who described a method for test equating without IRT. (In the Zu & Puhan research the context was more directly analogous to IRT equating – no judgements of difficulty were made, and all the items on both the 'old' and the 'new' test had been used before).

The main extra assumption needed here (apart from the obvious one that examinees on the grade boundary would have the same expected score as previous boundary examinees on the items judged/deemed to be identical) is that the grade boundaries were set in the 'correct' place on all previous versions of the component. The standard-maintaining procedures for A level examinations focus (rightly) on outcomes at the aggregate level. This means that anomalies and discrepancies can arise at unit/component level, particularly in assessments with a modular structure. That is, that the grade boundaries on units taken in January are not aligned with those taken in June (e.g., Black, 2008; Bramley, Dawson & Newton, 2014). However, this potential drawback would not apply to assessments with a more simple structure, and will be less relevant in future to A levels and GCSEs in England, where the reformed versions of both qualifications will be linear.

Table 5: Content of the questions for which an effectively identical counterpart was identified by Expert 2

June 2014 Question	*Max mark*	*Content of June 2014 question*	*Content of a previous question*
Q1bi	1	Antimony exists as a mixture of isotopes. What is meant by the term *isotopes*?	Tungsten has many isotopes. Explain what is meant by isotopes.
Q1biii	1	Complete the table below to show the atomic structure of ^{121}Sb. (Table with heading 'protons' 'neutrons' and 'electrons').	The mass number of one isotope of tungsten is 184. Complete the table below to show the atomic structure of this tungsten isotope. (Table with heading 'protons' 'neutrons' and 'electrons').
Q1ci	3	The relative atomic mass of antimony is 121.8. Define the term *relative atomic mass*.	Define the term *relative atomic mass*.
Q1dii	2	SbCl3 molecules are polar. Explain why.	Molecules of BF3 contain polar bonds, but the molecules are non-polar. Suggest an explanation for this difference.
Q2a	2	A compound used as a fertiliser has the following composition by mass: C, 20.00%; H, 6.67%; N, 46.67%; O, 26.66%. Calculate the empirical formula of this compound.	A compound containing magnesium, silicon and oxygen is also present in rock types in Italy. A sample of this compound weighing 5.27g was found to have the following composition by mass: Mg, 1.82g; Si, 1.05g; O, 2.40g. Calculate the empirical formula of the compound. Show your working.
Q4ai	2	H2O has hydrogen bonding. Complete the diagram below to show hydrogen bonding between the H2O molecule shown and one other H2O molecule. Include relevant dipoles and lone pairs. Label the hydrogen bond.	The solid lattice structure of ammonia, NH_3, contains hydrogen bonds. Draw a diagram to show hydrogen bonding between **two** molecules of NH_3 in a solid lattice. Include relevant dipoles and lone pairs.
Q4b	1	Draw a *'dot-and-cross'* diagram to show the bonding in CO2. Show outer electrons only.	Draw a *'dot-and-cross'* diagram to show the bonding in a molecule of CH_3Cl. Show **outer** electrons only.
Q5a	3	The Periodic Table is arranged in periods and groups. Elements in the Periodic Table show a periodic trend in atomic radius. State and explain the trend in atomic radius from Li to F. *In your answer you should use appropriate technical terms, spelled correctly*. (Answer space with one line for 'trend' and six lines for 'explanation').	Periodicity is a repeating pattern across different periods. First ionisation energy shows a trend across Period 2. The first ionisation energies of lithium, carbon and fluorine are shown in **Table 5.1** below. (Table giving the 3 values). Explain the trend across Period 2 shown in **Table 5.1**. *In your answer you should use appropriate technical terms, spelled correctly.*
Q5biii	1	A student adds a small volume of aqueous silver nitrate to an aqueous solution of bromide ions in a test-tube. The student then adds a similar volume of dilute aqueous ammonia to the same test-tube. Write an ionic equation for any precipitation reaction which occurs in the student's tests. Include state symbols.	A student was provided with an aqueous solution of calcium iodide. The student carried out a chemical test to show that the solution contained iodide ions. In this test, a precipitation reaction took place. Write an ionic equation, including state symbols, for the reaction that took place.
Q5cii	3	Under different conditions, chlorine reacts differently with aqueous sodium hydroxide. A disproportionation reaction takes place as shown below. (Chemical equation given.) State what is meant by disproportionation and show that disproportionation has taken place in this reaction.	The hydrides of Group 5 elements all exist as gases at room temperature. Phosphine gas, PH3, can be prepared by adding phosphorus, P4, to warm concentrated aqueous sodium hydroxide as shown in the equation below. (Chemical equation given.) Using oxidation numbers, explain why this is a disproportionation reaction.
Q6ai	1	Group 2 carbonates undergo thermal decomposition. Write the equation for the thermal decomposition of calcium carbonate. Include state symbols.	Magnesium carbonate, $MgCO_3$, is present in dolomite […] A student collected two equal-sized samples of dolomite. These samples were put into two labelled test-tubes, **A** and **B**. Tube **A** was heated until there was no further change in mass and was then allowed to cool. Tube **B** was left unheated. Write the equation for the action of heat on the magnesium carbonate present in tube **A**.

Unlike the first method, this method requires item level score data from the new test, so cannot be used before the test is taken (although the judgements of item similarity can of course be made in advance). However, the results of this study suggest it may be a more stable method in the sense of being less susceptible to differences among the judges, because the outcomes did not vary much when the criteria for identifying similar items were varied and different subsets of items were used to derive the boundary on the new test. Boundaries derived by this method were within 3 marks of the actual boundary in all cases, even with a very strict criterion for identifying pseudo-anchor items which identified only two items (worth 4 marks). Relaxing the criterion to include 20 marks worth (a third of the paper) of pseudo-anchor items produced boundaries only 1 mark away from the actual boundaries. Future work could explore whether it is better to use a few highly similar or identical questions as pseudo-anchors, or a larger number of less similar questions.

A further notable advantage of the second method is that it does not require the experts to make any judgements about mean scores of examinees, but just requires them to identify similar or identical questions. This should help to strengthen stakeholder confidence in the results by removing doubts about the ability of judges to make absolute (or indeed relative) judgements of difficulty. It also arguably allows the experts to give a more objective rationale for why they have deemed questions to be similar or identical, one which is more open to public scrutiny. For example, in a context such as in England where exam papers are published after they have 'gone live' the AO could publish the list of questions and their previous similar/identical counterparts that were used to derive the boundaries. Such a list is provided for this study in Table 5.

A limitation of this study is that it only involved two expert judges. This was necessary to meet the strict security conditions surrounding research in a 'live' setting. However, the expertise of the judges was as high as it would be possible to achieve, involving as it did the most experienced and senior examiners involved in setting the examination. It is an open question whether widening the pool of judges would improve the estimates (by reducing random error) or make them worse (by introducing bias and/or random error from relative lack of expertise).

Further work is needed to determine how well the findings from these two methods will generalise to other assessments than the one studied here. It seems reasonable to expect that judgements about question difficulty or similarity are better suited to exams consisting of relatively objective shorter answer questions where 'question difficulty' is a more tangible concept, and where it may be easier to define the knowledge and skills required to answer a question. The component used in the study reported here had a very large entry with hundreds of examinees on each mark point in the score distribution. More technical work could focus on the numbers of examinees needed to allow satisfactory estimates of the EICCs and experiment with varying the smoothing parameter to see what effect it has on the results.

In conclusion, both methods show promise for use in operational standard-maintaining procedures in contexts where tests are constructed to the same general specifications, but there is no possibility for pre-testing or re-use of items. In the context of GCSEs and A levels in England, these methods could provide a good source of relatively independent evidence about the difficulty of the questions, which could complement the existing evidence about the ability of the examinees and the quality of their work in the examination.

Acknowledgements

We would like to thank Gwen Low, formerly of OCR, and the two examiners who acted as expert judges for their support with this research.

References

Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating item difficulty with comparative judgments. *ETS Research Report Series, 2014*(2), 1–8. Available online at: doi: 10.1002/ets2.12042

Baird, J.-A., Cresswell, M., & Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education*, *15*(2), 213–229. Available online at: doi: 10.1080/026715200402506

Bejar, I. I. (1983). Subject Matter Experts' Assessment of Item Statistics. *Applied Psychological Measurement*, *7*(3), 303–310. Available online at: doi: 10.1177/014662168300700306

Benton, T. & Bramley, T. (2015). *The use of evidence in setting and maintaining standards in GCSEs and A levels.* Cambridge: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/Images/204310-maintaining-standards-discussion-paper-tom-benton-and-tom-bramley.pdf

Benton, T., & Lin, Y. (2011). *Investigating the relationship between A level results and prior attainment at GCSE*. Coventry: Ofqual.

Black, B. (2008). *Using an adapted rank-ordering method to investigate January versus June awarding standards*. Paper presented at the Fourth Biennial EARLI/Northumbria Assessment Conference, Berlin, Germany. Available online at: http://www.cambridgeassessment.org.uk/images/109767-using-an-adapted-rank-ordering-method-to-investigate-january-versus-june-awarding-standards.pdf

Bramley, T. (2010). 'Key discriminators' and the use of item level data in awarding. *Research Matters: A Cambridge Assessment Publication*, *9*, 32–38. Available online at: http://www.cambridgeassessment.org.uk/Images/109986-research-matters-09-january-2010-.pdf

Bramley, T., Dawson, A., & Newton, P. E. (2014). *On the limits of linking: experiences from England*. Paper presented at the 76th annual meeting of the National Council on Measurement in Education (NCME), Philadelphia, PA.

Brandon, P. R. (2004). Conclusions about Frequently Studied Modified Angoff Standard-Setting Topics. *Applied Measurement in Education*, *17*(1), 59–88. Available online at: doi: 10.1207/s15324818ame1701_4

Clauser, J. C., Margolis, M. J., & Clauser, B. E. (2014). An Examination of the Replicability of Angoff Standard Setting Results within a Generalizability Theory Framework. *Journal of Educational Measurement*, *51*(2), 127–140. Available online at: doi: 10.1111/jedm.12038

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education*, *25*(3), 271–284. Available online at: doi: 10.1080/02671522.2010.498143

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, *15*(4), 237–261. Available online at: doi: 10.1111/j.1745-3984.1978.tb00072.x

Hambleton, R. K., & Jirka, S. J. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp.399–420). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Humphry, S., Heldsinger, S., & Andrich, D. (2013). Requiring a consistent unit of scale between the responses of students and judges in standard setting. *Applied Measurement in Education*, *27*(1), 1–18. Available online at: doi: 10.1080/08957347.2014.859492

Impara, J. C., & Plake, B. S. (1998). Teachers' Ability to Estimate Item Difficulty: A Test of the Assumptions in the Angoff Standard Setting Method. *Journal of Educational Measurement*, *35*(1), 69–81. Available online at: doi: 10.1111/j.1745-3984.1998.tb00528.x

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

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

Ofqual. (2011). *GCSE, GCE, Principal Learning and Project Code of Practice*. Coventry: Ofqual. Available online at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/371268/2011-05-27-code-of-practice.pdf

Taylor, M. (2013). *GCSE predictions using mean Key Stage 2 level as the measure of prior attainment*. Report to Joint Council on Qualifications (JCQ) Standards and Technical Advisory Group (STAG). Revised 26/06/13.

Zu, J., & Puhan, G. (2014). Preequating with empirical item characteristic curves: an observed-score preequating method. *Journal of Educational Measurement*, *51*(3), 281–300. Available online at: doi: 10.1111/jedm.12047

Research News

Karen Barden Research Division

Conferences and seminars

European Conference on Educational Research (ECER)

The ECER conference took place in Budapest, Hungary in September, under the theme of *Education and Transition – Contributions from Educational Research*. Nadir Zanini, Research Division, presented a paper on *The importance of teaching styles and curriculum in Mathematics: Evidence from TIMSS 2011*. The paper was co-authored with Tom Benton, Research Division.

Simon Child, OCR, presented a paper co-authored with Research Division colleagues Prerna Carroll and Ellie Darlington on *The role of assessment in facilitating student transition to 'active' citizenship*.

Royal Statistical Society (RSS)

The RSS 2015 Annual Conference took place in Exeter in September. Now in its 23rd year, the RSS conference has gained prestige for its focus on current statistical issues, how it fosters the exchange of ideas and information, and the quality of its speakers. Tom Benton, Research Division, presented a paper on *How statistics determine examination results in England*.

British Educational Research Association (BERA)

Held in September at Queen's University, Belfast, Northern Ireland, the BERA Annual Conference was an opportunity to develop new research ideas, and to build new research relationships within the research education community. Based on work undertaken by the Research Division, Cambridge Assessment colleagues presented the following papers:

Carmen Vidal Rodeiro, Research Division: *An investigation into the numbers and characteristics of candidates with incomplete entries at AS/A level*.

Simon Child, OCR, Ellie Darlington and Tim Gill, Research Division: *An investigation of the motivations underpinning student and teacher topic choice in History qualifications*.

Jessica Bowyer (née Munro), Research Division: *The assessment of creativity and innovation in Design and Technology*.

Martin Johnson, Research Division: *Reading between the lines: exploring the characteristics of feedback that support examiners' professional knowledge building*.

Tim Gill, Carmen Vidal Rodeiro and Nadir Zanini, Research Division: *Students choices in Higher Education*.

Jackie Greatorex, Lucy Chambers, Filio Constantinou and Jo Ireland, Research Division: *Piloting a method for comparing examination question paper demands*.

Jackie Greatorex, Tom Sutch, Jessica Bowyer, Karen Dunn, Research Division, and Magda Werno, Cambridge International Examinations: *Investigating a new method for standardising essay marking using levels-based mark schemes*.

Victoria Crisp, Research Division: *Validity and comparability of assessment: how do these concepts relate?*

Magda Werno, Cambridge International Examinations, Frances Wilson, OCR, and Prerna Carroll, Research Division: *Translation in the reformed ancient languages GCSEs*.

Gender differences – the impact of secondary schooling – boys or girls, who's winning?

A Cambridge Assessment conference on 'Gender differences' took place in London in October. The conference brought together more than 600 experts from within the education and assessment community both at the conference and online, with over 30 countries represented. The audience heard from speakers from around the world who unpacked the complex range of issues that surround gender differences in secondary education and how they might be tackled to attempt to remove, or at least start to reduce, the gap between girls and boys. Presentations included the following papers:

Tim Oates, Assessment, Research & Development: *An analysis of the gender divide – from primary school to workforce*.

Tom Benton, Research Division: *Attitudes to learning – questioning the PISA data*.

Tom Bramley, Carmen Vidal Rodeiro and Sylvia Vitello, Research Division: *Gender differences at GCSE*.

Agnieszka Walczak and Ardeshir Geranpayeh, Cambridge English Language Assessment: *The Gender Gap in English Language Proficiency? Insights from a Test of Academic English*.

Further details of the conference, videos of the proceedings and additional resources can be found on our website at: http://www.cambridgeassessment.org.uk/events/gender-differences-conference-2015/