Summer 2018
No. 1

INSIDE ASSESSMENT

Research news & comment

The Standards issue

Compare and contrast
Comparative judgement in awarding

Bayesian approaches
Weighing statistics and values

Comparable outcomes
The latest developments

PLUS
International standard setting
Meet the researchers
From the archives

Centre for Education Research and Practice (CERP)
cerp.org.uk

AQA
Realising potential

INSIDE ASSESSMENT

Research news & comment

Summer 2018 No. 1
The Standards issue

News and findings from AQA's Centre for Education Research and Practice (CERP)

Edited by
Emma Armitage, Lena Gray, Ruth Johnson
Lindsay Simmonds and Steve Wooding

Contributors
Cesare Aloisi, Yaw Bimpeh, Simon Eason,
Ben Jones, Kate Kelly, Lesley Meyer,
Alex Scharaschkin and Martin Taylor

Inside assessment
Published by AQA

Further articles and research papers are available to download at cerp.org.uk
Please send editorial correspondence to cerp@aqa.org.uk

Edited, designed and produced by Claire Jackson
Images © Shutterstock unless otherwise stated
Printed by Optichrome

CONTENTS

SUMMER 2018

From Alex Scharaschkin, AQA Executive Director of Research and Compliance

DEFINING AND ASSESSING standards of attainment in education is challenging. Attainment combines knowledge, skills and understanding; these attributes are assessed in a variety of ways. Different individuals place different values on these aspects, so educational standards are effectively social constructs that reflect what is regarded as valuable in a particular context.

Once attainment standards have been set, they must be assessed consistently over the lifetime of a qualification. Standards maintenance is particularly vital in high-stakes examinations that affect young peoples' life chances. In England, this relates to the General Certificate of Secondary Education (GCSE), which students take at the age of 16, and a further subject-specific qualification, the A-level, taken thereafter. (For contextual information, see pp. 46–47.)

In the system in England, the need for fairness means that comparability extends to standards set by different awarding organisations, in different subjects and in different years. AQA has a significant body of research that has focused on the methodological aspects of comparability studies, drawing on both the statistical and judgemental elements of the process. Recent changes in policy have meant that researchers have latterly been involved with establishing and describing the standards of a new grading scale, and ensuring comparability with past and present systems.

Initial techniques for standards maintenance comprised the so-called delta analysis (monitoring entry patterns across awarding bodies and between years); common centres analysis (looking at results for centres entering candidates for a subject in successive years); and

subject pairs (candidates entering two subjects were expected to obtain similar results in those subjects). Major AQA research projects looked into the use of judgement in awarding, resulting in the much-cited 'Good & Cresswell effect', which can be summarised as the tendency of examiners to compensate insufficiently for variation in question paper/mark scheme demand when deciding on grade boundaries. We are pleased to be able to publish an archive paper on this topic here ('Setting comparable standards on examination papers of differing difficulty' by Mike Cresswell and Frances Good, pp. 35–45.)

Technological and theoretical development has allowed standards research to take new directions. AQA's current research programme – carried out within the Centre for Education Research and Practice (CERP) – draws on the work of its past research units (including within the Joint Matriculation Board and the Associated Examining Board) with its gaze firmly on the future. The articles within this first volume of our new publication, *Inside assessment*, give a flavour of activity.

From comparative judgement (pp. 18–19), to Bayesian approaches to standard setting (pp. 20–25) and subject-specific case studies (pp. 14–17), our researchers tackle a range of topics, using a variety of techniques. But while the methods and the research continue to evolve, the aim remains the same: fairness to students. ■

Alex Scharaschkin became Director of the Centre for Education Research and Practice (CERP) in July 2014, and was appointed Executive Director of Research and Compliance in November 2015. He was previously Director for Regulation, Consumers and Competition at the National Audit Office (NAO), where he led the NAO's work examining the government's use of markets in the private and public sectors. Alex has a background in assessment research: he was Principal Officer for Statistical Analysis at the Qualifications and Curriculum Authority, and held research posts at the Associated Examining Board and the Institute of Education, University College London. Alex also served as a member of CERP's advisory group for four years and is currently Executive Secretary of AEA-Europe, a leading association for educational assessment researchers and practitioners across Europe.

What *are* standards?

Measuring performance

If we want to know the length of something, we can measure it against a ruler.

There is no definitive measure of a GCSE or A-level against which we can compare a question paper or the quality of a student's performance.

Standards are defined by subject content and assessment design experts.

Standards are multidimensional

Content standards = the demand of the specification

Assessment standards = the demand of an exam paper

Performance standards = the quality of students' responses

How do we define standards?

Sociological perspective
Grades represent a value judgement about the quality of students' work made by experts who are members of a community of practice.

Strong criterion referencing
Grades represent that a student possesses specific knowledge, skills and understanding.

Weak criterion referencing
Grades represent the general quality of students' work rather than mastery of specific knowledge, skills and understanding.

Catch-all definition
Grades represent a student's position in the rank order after all factors predictive of exam success have been controlled for. They do not reveal anything about the quality of students' work.

Simple cohort referencing
Grades represent a student's position in the rank order (determined by marks achieved) of his or her cohort. They do not communicate anything about the quality of student's work.

REFERENCES

Cresswell, M. J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues* (pp. 57–84). London: John Wiley & Sons Ltd.

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

Different paths

Emma Armitage rounds up past and present views on the role of expert judgement in standard setting – and highlights burgeoning areas of research

STANDARD SETTING IS the process of establishing grade boundaries on assessments in order to separate students into two (pass/fail) or more (GCSE grades 9 to 1) distinct categories of performance (Cizek & Bunch, 2007). In England, this process involves asking expert judges – often current or former teachers – to review students' completed exam papers and evaluate the quality of their responses. For example, experts might be asked to consider: 'Is this script worthy of a grade 5?' This is a challenging task, particularly in its current incarnation, because it involves balancing a number of sources of evidence. It also requires judges to distinguish work worthy of a particular grade from work unworthy of that grade within a very narrow range of marks (often just five).

In recent years, less weight has been placed on expert judgement. There is increasing availability of statistical evidence that can be used to set grade boundaries. However, in situations where the statistical evidence is weak, a heavier reliance on expert judgement is still necessary to ensure defensible grade boundaries (Jones, 2015).

Retaining the involvement of subject experts is important; their scrutiny of students' work is key to external stakeholders since it allows grades to be interpreted as representations of what candidates know and can do at a given level of attainment. However, the role played by expert judgement in standard setting needs to be reshaped if it is to provide a meaningful contribution. The first step in that process is to review the existing problems with using expert judgement in the context of standard setting in England; these problems can be categorised according to their source:

1) The judges

Judges are expected to be familiar with how students might perform on an exam and to be experts in the exam content. For this reason, the judges are usually teachers. Even so, unique experiences with students of varying abilities and different levels of content knowledge can influence evaluations of students' work. Judges who teach high-performing classes have been found

to recommend higher grade boundaries than judges with lower-performing classes (Nasstrom & Nystrom, 2008). Similarly, judges sometimes set more severe standards on questions they know the correct answers to, because they also expect students to answer correctly. (Chang, Dziuban, Hynes, & Olson, 1996).

2) The work under scrutiny

Judges are required to consider several students' work on a single mark. However, those pieces of work may look quite different. There are multiple ways to achieve the same total mark: one student could score average marks on every question (consistent), while another student could accrue very high marks on a small number of questions but low marks on others (inconsistent). The manner in which students accumulate marks has been shown to influence judges' evaluations of their grade worthiness. Consistent scripts are more often considered gradeworthy than inconsistent scripts. Moreover, judges report finding it more difficult to judge inconsistent scripts, most likely because they activate different perceptions of a candidate's ability (Scharaschkin & Baird, 2000).

3) The evidence provided to assist judges in making judgements

Judges are given a variety of information to help them evaluate the quality of students' work; this includes the total marks students achieved and statistically recommended boundaries (SRBs) – the marks at which statistical predictions suggest grade boundaries should be set. Research has demonstrated that the evaluations made by judges are at least somewhat reliant on these benchmarks. When students' marks are removed from their examination papers, judges are not very good at rank-ordering the papers in terms of quality; however, when the marks are present, judgements of quality generally correlate well with the rank order of marks (Baird & Dhillon, 2005). In instances where judges have been given incorrect SRBs, very few judges spot the error, and, where they do, they typically fail to adequately compensate for it (Stringer, 2012).

Experts' scrutiny of students' work is key to external stakeholders since it allows grades to be interpreted as representations of what candidates know and can do at a given level of attainment

4) The situation in which the judgements are made

Most often, standard setting takes place in a group setting. Judges make individual evaluations of the quality of students' work that are then shared with the group. Discussion of the individual judgements tends to result in consensus. However, reaching a consensus does not mean the final judgement is accurate or

valid; its validity rests on how consensus is reached. Discussion allows the exchange of rational arguments regarding the appropriate location of a grade boundary, but it can also foster group dynamics. These dynamics may influence members' judgements or allow individual members to assert dominance over the group. Research evidence is split on this topic. It has been discovered that if an individual is presented with a consensus opinion, they may feel pressure to conform (Murphy et al., 1995). It has also been found that dominant group members have a disproportionate influence on the recommended grade boundary (Brennan & Lockwood, 1980). However, other work has found no evidence that dominant group members unequally influence grade boundaries (Williams, Klamen, & McGaghie, 2003).

It should be evident from this brief overview that deciding what should legitimately influence experts' judgements of quality or how the judgement process should be organised is far from clear-cut: performance standards do not exist 'independently of human opinions and values' (Shepard, 1979, p. 62), and therefore there is no objective right answer as to where grade boundaries should be positioned or how they should be derived. Nonetheless, this exercise of collating what we know about the role of expert judgement should help provide clarity.

In the longer term, it may be possible to redefine the judgement task in order to more clearly delineate the factors that judgements should be based on, capitalising on what judges can do well, while protecting against what they cannot. ▣

Emma Armitage joined CERP in May 2015, having completed a PhD in Psychology at Lancaster University. She also holds a BSc in Psychology and an MSc in Developmental Disorders. Emma's doctoral research explored three- to eight-year olds' understanding of pictures as symbols, with a specific focus on emerging knowledge of different picture mediums. Her paper on this topic was published in the journal *Developmental Psychology*.

REFERENCES

Baird, J., & Dhillon, D. (2005). *Qualitative expert judgements on examination standards: valid, but inexact.* Manchester: AQA Centre for Education Research and Practice. Internal report.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using Generalizability Theory. *Applied Psychological Measurement*, *4*, 219–240.

Chang, L., Dziuban, C. D., Hynes, M. C., & Olson, A. H. (1996). Does a standard reflect minimal competency of examinees or judge competency? *Applied Measurement in Education*, *9*, 161–173. doi: 10.1207/s15324818ame0902_5

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Jones, B. E. (2015). *A better way to award?* Manchester: AQA Centre for Education Research and Practice. Internal report.

Murphy, R., Burke, P., Content, S., Frearson, M., Gillespie, J., Hadfield, M., Rainbow, R., Wallis, J., & Wilmut, J. (1995). *The reliability of assessment of NVQs*. Report to the National Council for Vocational Qualifications, School of Education, University of Nottingham.

Nasstrom, G., & Nystrom, P. (2008). A comparison of two different methods for setting performance standards for a test with constructed-response items. *Practical Assessment, Research & Evaluation*, *13*, 1–12.

Scharaschkin, A., & Baird, J. (2000). The effects of consistency of performance on A level examiners' judgements of standards. *British Educational Research Journal*, *26*, 343–357.

Shepard, R. N. (1979). Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based education* (pp. 59–71). Washington, DC: National Council on Measurement in Education.

Stringer, N. S. (2012). Setting and maintaining GCSE and GCE grading standards: the case for contextualised cohort-referencing. *Research Papers in Education*, *27*, 535–554. doi:10/1080/02671522.2011.580364

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Special article: Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine*, *15*, 270–292. doi: 10.1207/S15328015TLM1504_11

Balancing act

Standards in modern foreign languages (MFL) are under scrutiny, due to a supposed impact of native speakers on prediction techniques, and the increasing relative proportion of students with a 'comparative advantage'. **Cesare Aloisi** details how researchers are unpicking the arguments, asking further questions and helping to develop solutions

MAINTAINING STANDARDS means ensuring that students displaying similar levels of ability in a subject receive the same final grade, even if examinations differ across subjects, and over time. In England, standards are maintained using two sources of evidence: experienced examiners' judgement and statistical data, as outlined on pp. 46–47.

Standards in modern foreign languages (MFL), which are optional at GCSE and A-level (see p. 46), have been the subject of recent attention, both in the political arena and within the research community. There have been suggestions that standards in MFL are at risk (e.g. WJEC: see Castle-Herbert et al., 2017; Ofqual: see Taylor & Zanini, 2017). Two main arguments have been posited. Firstly, that the presence of native speakers makes prediction matrices – an algorithm used to summarise the proportion of students in each mean GCSE decile who achieved each A-level grade (A*–E) in a reference year and for a given subject – inaccurate. This is because the correlation between the GCSE mean decile and the GCSE results in MFL is weaker than that for other subjects. Therefore, using this standard relationship to decide on cut-scores might not work for MFL as it does for other subjects.

Secondly, the continuous decrease in language take-up at GCSE and A-level might be increasing the relative proportion of native speakers or students with a 'comparative advantage' (Castle-Herbert et al., 2017). If the student composition has changed from the reference cohort, following the prediction matrix may result in skewed cut-scores: at the upper end of the score distribution, more students would achieve high grades. However, adjusting the cut-scores to bring the percentages in line with previous years means that even students who did very well may miss a top grade by one or two

What does it mean to be a native speaker? The everyday meaning does not withstand scientific scrutiny – the apparent straightforwardness of the concept masks several difficulties

points. This would be unfair to both native speakers (similar abilities might result in different grades) and to the rest of the cohort (it would become artificially difficult to achieve a good grade in the subject). Some solutions to monitor and tackle these issues have been proposed. These include comparative progression analysis (Newton, He, & Black, 2017), expanded prediction matrices (Castle-Herbert et al., 2017) and generalised boosting models (Benton, 2015).

The first two approaches consider GCSE MFL achievement, not just mean decile, to calculate A-level predictions; the third approach uses machine learning algorithms to improve such predictions. CERP's current research involves further investigation of these approaches. The overall aim is to contribute to the development of new techniques to advance the reliability of standards.

As part of our analysis, we highlight the following points regarding the argument that MFL standards are at risk:

- Previous analyses do not clarify with sufficient theoretical detail what it means to be a native speaker. The everyday meaning does not withstand scientific scrutiny – the apparent straightforwardness of the concept masks several difficulties.

  For example, the first language one learns might not be that mastered at the age of 16, and being born in a multilingual family does not guarantee proficiency in the minority language. In fact, children growing up in a household where a minority language is spoken alongside English are likely not to speak it (De Houwer, 2017). This situation occurs even more frequently among second- or third-generation British children: in many cases, these children may be able to say a few sentences in the heritage language with a good accent, but not much more. Therefore, they might not

necessarily achieve top marks at GCSE or A-level.

- Comparative advantage may affect a variety of subjects – not just MFL. It is well known that family background influences final achievement, so each student may have a relative advantage over other students in a subject (or across all subjects, if one considers socioeconomic advantage). There are various subjects for which access to out-of-school educational opportunities might provide an advantage – such as Music, Art & Design, Computer Science or Religious Education – yet availability does not necessarily translate into better performance.

Ofqual recognises that prediction matrices appear to have worked appropriately in MFL, as they have across all other subjects. For instance, 'none of A level French, German and Spanish had outcomes that exceeded reporting tolerances in June 2015 or 2016' (Taylor & Zanini, 2017, p. 58). Benton (2015) found that MFL predictions were on a par with the accuracy reported for Art & Design or Chemistry, and lower than that for Mathematics. There have been no calls to review the standard-setting procedures to account for the hypothetical presence of people with an innate talent – and therefore a comparative advantage – in other subjects.

We are contributing to the refinement and development of techniques to set standards across subjects and over time, with a view to advancing the field regardless of any decision that might be made on MFL. We are doing this by:

- Extending Ofqual's work on comparative progression analysis, to understand whether GCSE to A-level progression patterns vary by language.

- Extending and updating WJEC's work to include all languages and the 2015-2017 trends.

- Supporting descriptive analyses with multilevel logistic models to detect statistical anomalies after controlling for mean GCSE grade, gender and prior achievement in the subject (if candidates sat the corresponding GCSE).

- Complementing these analyses with a more thorough investigation on the estimated proportion of 'native speakers' (subject to some constraints in the definition) in MFL and their achievement, by using data from the National Pupil Database. ▣

REFERENCES

Benton, T. (2015). Can we do better than using 'mean GCSE grade' to predict future outcomes? An evaluation of generalised boosting models. *Oxford Review of Education*, *41*, 587–607. doi:10.1080/03054985.2015.1074563

Castle-Herbert, A., Evans, A., Maziarz, J., & Morgan, P. (2017, November). *Changing abilities and comparable outcomes: UK examinations of French, Spanish and German*. Poster presented at AEA-Europe conference, Prague.

De Houwer, A. (2017). Minority language parenting in Europe and children's well-being. In N. J. Cabrera & B. Leyendecker (Eds.), *Handbook on positive development of minority children and youth* (pp. 231–246). Cham: Springer. doi:10.1007/978-3-319-43645-6

Newton, P. E., He, Q., & Black, B. (2017). *Progression from GCSE to A level: Comparative progression analysis as a new approach to investigating inter-subject comparability*. Coventry: Ofqual Office of Qualifications and Examinations Regulation.

Taylor, R., & Zanini, N. (2017). *Native speakers in A level modern foreign languages.* Coventry: Ofqual Office of Qualifications and Examinations Regulation.

Cesare Aloisi joined CERP in July 2017. Prior to this, he worked at the University of Reading on a project analysing student learning trajectories, critical thinking, engagement and wellbeing. He has a PhD in Education from the University of Durham and an MA in Educational Assessment from University College London. Cesare is a former language teacher and has an interest in large-scale assessments, multilevel modelling, early childhood education and social justice issues.

Compare & contrast

Statistical data and examiner judgement are both used to determine the position of grade boundaries, but over time, data has come to dominate the process. **Kate Kelly** outlines how comparative judgement may redress the balance

SETTING AND MAINTAINING examination standards is a core part of AQA's work, a reality emphasised throughout this publication. In England, standard setting is achieved by adjusting the cut-scores (grade boundaries) of the question papers, with the aim of cancelling out any differences that may have arisen between cohorts due to differing paper difficulties. This process is known as awarding. Typically, two sources of evidence are used to decide the position of grade boundaries: statistical data and examiner judgement.

Over time, the use of statistical data has come to dominate the process. This is largely due to research that has highlighted the limitations of examiners' ability to make the fine judgements required. Nevertheless, examiner judgement has an important role to play. Examinations rely on public trust to retain their currency, and it is unclear whether an entirely statistical standard would be widely accepted. Consequently, the use of examiner judgement has become a pressing issue for many awarding organisations in England. An increasingly popular suggestion is to use comparative judgement.

A comparative judgement is one in which two or more stimuli are judged in relation to each other on the basis of some criteria. In standard setting, this would mean asking an examiner to evaluate which of two student responses is better. This sits in contrast to what is usually termed 'absolute judgement', in which the examiner is asked to evaluate the quality of a single student response.

When multiple comparative judgements are made, by multiple judges, they can be aggregated together to produce a rank order of responses, usually with a high level of reliability. Depending on the exercise, this rank order of responses can be used in a number of ways, including: to replace marking, to compare performance across cohorts, to compare performance over time, and, potentially, to set grade boundaries.

The method has been found to yield consistently high estimates of reliability – often higher than can be achieved using conventional approaches

Methods based on comparative judgement have gained a great deal of popularity in recent years, for a number of reasons. It is clear that comparative judgement overcomes a number of the issues associated with current methods of using judgements in awarding. For instance, eliciting relative judgements means that differences between examiners in terms of leniency and severity are eradicated. The method has also been found to yield consistently high estimates of reliability – often higher than can be achieved using conventional approaches. The main point in favour is that relative judgements are also believed to be cognitively easier than absolute judgements. Comparative judgement, it is argued, capitalises on humans' natural approach to such tasks.

However, the extent to which this argument holds is unclear. Those advocating for comparative judgement have yet to marshal the psychological evidence to support this claim – nor has its relevance to judgement in this context been fully explored. Moreover, the validity evidence for comparative judgement is by no means clear-cut.

Of course, such criticisms apply also to the current approach to eliciting examiner judgements, and the psychological aspects of how examiners evaluate performance quality are under-researched in general. As such, any method of awarding that relies on judgement alone – comparative or otherwise – is unlikely to be satisfactory for such high-stakes examinations as GCSEs or A-levels.

Nevertheless, with further research, comparative judgement could make a valuable contribution to the current repertoire of standard-setting and maintenance techniques. ■

Kate Kelly joined the Centre for Education Research and Practice (CERP) in June 2010 after completing a BSc (Hons) in Psychology at the University of Bath, which included a year spent with CERP as a placement student. Kate has an interest in novel methods of test equating using comparative judgement difficulty estimates and item facilities. She is currently working towards a PhD on the potential of comparative judgement for improving grading decisions.

Bayesian statistics: a bluffer's guide

Alex Scharaschkin sums up the Bayesian approach to standard setting and maintenance

THE NEED TO COMBINE evidence of varying degrees of reliability is a common issue in many sectors, including: audit, medicine, law, and risk analysis. There are a number of ways of addressing the challenge, but the so-called Bayesian approach is increasingly popular. Recent CERP studies have investigated how this technique could be used to combine statistical information and expert judgement when setting grade boundaries (see pp. 46–47).

The first key idea in Bayesian statistics is that the probability of a particular assertion represents our degree of belief, or confidence, in that assertion. More particularly, we can represent our confidence in the position of, say, the fairest grade C boundary mark for an examination, by a probability distribution. The distribution gives us both what we expect the value to be (the 'expected value', or average), and how confident we are in that expectation (the 'variance', or degree of spread around the average). When we come to set a grade boundary in a subject, we have prior information in the form of the statistical prediction. If we are confident in the statistics, then we will have an expected value for, say, the grade C boundary mark. The prior attainment-based statistics might suggest, for example, that it is highly likely that the C boundary should be set at 60 marks. They might also suggest that, while 60 is the most likely mark, it is also possible, though less likely, that it could be 59 or 61, but that it is highly unlikely (practically zero probability) that it could be lower than 59 or higher than 61. In this case, we have a so-called prior distribution for the position of the C boundary mark that has an expected value of 60, and a small variance around that expected value. In another subject, the statistics might be less reliable. We might only have a small number of matched candidates (for whom we have prior attainment information), and more uncertainty in our predictions. In such a case, we would have a prior distribution with a larger variance, reflecting less certainty about the expected position of the grade boundary.

The second key idea in Bayesian statistics is that in situations where we have a prior belief or degree of certainty about a proposition (such as the

position of the grade C boundary mark in an examination), we can update this prior view in the light of further information that comes into our possession. Bayes' rule tells us how to do this, to obtain a so-called posterior distribution that reflects both our prior beliefs, and the new data. If the new data we are given, or observe, happen to match our prior beliefs quite closely, then our posterior distribution will not be much different from our prior. If, on the other hand, the new data is relatively discordant with our prior beliefs, then our posterior distribution is likely to shift somewhat towards what the new data suggests. The amount of 'shift' depends in part on how strong our prior beliefs are (i.e. the variance in our prior).

The more certain we are of a priori, the more evidence we need to change our views, and the less likely we are to accommodate discordant evidence fully. How much we shift also depends on how far away from our expectations the new data are. Bayes' rule updates a given prior, to generate a posterior distribution, using what is known as the 'likelihood function'—the conditional probability, given the prior distribution, that the new data would actually be distributed in the way we observe them to be.

In the context of grade boundary setting, the prior distribution is derived from the statistical information. The likelihood function is then derived from the empirical information on the awarders' judgements. There are a number of ways that this information (i.e. the judged grade for each script that was looked at by a member of the awarding committee) can be elicited. CERP research uses the tick chart records from awarding meetings as representing the judgemental data.

These records contain each awarding committee member's decision regarding recommended grade boundaries. After the awarding committee has scrutinised scripts on a range of marks for the grades at which recommendations are required (called the key boundaries: A, C and F for GCSE examinations; A and E for GCE), each is asked, in turn, to give his/her decisions on the scripts he/she has considered. These decisions are recorded on a chart. For each script they have seen, each awarder is asked to give his/her decision as to whether it is 'in the grade category' or 'outside the grade category' of the grade in question. When the awarder is unsure whether the script is worthy of the grade in question, it is recorded on the chart as '?'. The chart is generally referred to as the 'tick chart'.

It is possible to then use Bayes' rule to derive a formula for weighting the judgemental and statistical evidence in accordance with the relative confidence we have in each, to produce a grade boundary mark that reflects both of these sources of evidence. ▣

Intelligent integration

Current awarding procedures rely on a combination of statistical information and expert judgement, but what if there were a new way to determine the weighting? **Yaw Bimpeh** describes a Bayesian approach to standards maintenance

IN THE CONTEXT OF HIGH-STAKES exams in England and Wales (GCSEs and A-levels), the process of setting performance standards or cut-scores is generally referred to as 'awarding' (as explained on pp. 46–47). Recent CERP studies have investigated possible uses of a Bayesian approach to modify and build on the current awarding procedure to meet the challenge of combining statistical information and expert judgement when setting grade boundaries.

While Bayesian statistics can be somewhat daunting, the method referred to here is relatively straightforward. The aim is to combine the two sources of evidence – statistical and judgemental – in a more empirical manner, as opposed to ad hoc. The statistical information takes the form of percentages of candidates we predict will get given grades, based upon their prior attainment. We use this information to work out the grade boundaries that would match this prediction. The 'judgement' information is similar – the experts recommend grade boundaries based on how candidate performance in the current year matches up with the previous year.

So, we are faced with two sets of recommended grade boundaries, based upon two sources of information, which we want to integrate. In order to do this, for each of the two sets of recommended boundaries, we can work out the proportion of candidates that would obtain each grade. An example of this can be found below:

	A*	A	B	C	D	E	U
Statistical boundary recommendations	71	65	59	54	49	44	–
Proportion of candidates	0.017	0.084	0.223	0.455	0.658	0.823	1.000
Judgemental boundary recommendations	71	64	57	51	45	39	–
Proportion of candidates	0.019	0.112	0.336	0.599	0.812	0.920	1.000

The resampling procedure lets us establish confidence intervals for our expected proportions. These give us a range of proportions, between which lies the 'true' value

The next step is to combine the two sources of information. Statistical and judgemental information is derived from data with different sample sizes. The estimates from the larger data will be more precise than the estimates from the smaller data. It is reasonable to give more weight to the more precise estimates when combining the statistical and judgemental information. This is achieved in a fairly straightforward way – by obtaining a weighted average of the two sets of proportions. Hence, the weight given to each source of information is key: it has a huge impact on the resulting combination of sources of information.

Unfortunately, working out numerically equivalent weights for the two sources is not an easy task. For the statistical information, we can use the number of matched candidates who had prior attainment information that was used to generate the prediction. Choosing a number for the judgement weighting is more difficult. We used the number of candidates who achieved marks within the range of marks scrutinised during the judgement exercise. This weighted average produces a combined set of proportions indicating the proportion of candidates we expect to achieve each grade. However, we can't solely use this combined proportion as the output of the process. Both judgemental and statistical information have a degree of uncertainty associated with them – and this doesn't disappear when we combine the two with a simple weighted average.

In order to try to account for this uncertainty, we sample and resample the weighted proportions thousands of times, giving us a final Bayesian recommendation for the proportion of candidates we expect should achieve each grade. Crucially, this sampling procedure also lets us establish confidence intervals for our expected proportions. These give us a range of proportions, between which lies the 'true' value (to a given degree of certainty, usually 95 per cent).

Finally, the Bayesian recommendation and these upper and lower confidence intervals – currently in the form of proportions – are translated back into grade boundaries by working out what mark would produce that proportion of candidates getting each grade. An example of this is given in the table overleaf:

		A*	A	B	C	D	E	U
Upper bound	Proportion of candidates		0.074				0.814	
	Boundary recommendation	72	66	60	55	50	45	0
(Sampled) Bayesian approach	Proportion of candidates	**0.018**	**0.092**	**0.256**	**0.497**	**0.703**	**0.851**	**1.000**
	Boundary recommendation	**71**	**65**	**59**	**54**	**48**	**43**	**0**
Lower bound	Proportion of candidates		0.111				0.889	
	Boundary recommendation	70	64	58	52	46	41	0

The recommended boundary positions in bold type are those that the Bayesian approach deems appropriate, based on an empirical combination of statistical and judgemental evidence. The lower and upper bounds indicate the range of possible values that the Bayesian method suggests the boundary should fall within, if the most likely position is not considered desirable. ▣

With thanks to Ben Smith for his input

REFERENCES

Baird, J., Cresswell, M. J., & Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education*, *15*, 213–229.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer Verlag.

Bramley, T. (2005). Accessibility, easiness and standards. *Educational Research*, *47*, 251–261.

Casella, G., & George, E. (1992) Explaining the Gibbs sampler. *The American Statistician*, *46*, 167–174.

Christie, T., & Forrest, G. M. (1981). *Defining public examination standards*. Schools Council Research Studies. London: Macmillan Education.

Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice*, *15*, 13–21.

Cizek, G. J. (2001). Conjectures on the rise and fall of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3–17). Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Congdon, P. (2005). Bayesian models for categorical data. Chichester: Wiley.

Cresswell, M. J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, Developments and Statistical Issues*. Chichester: Wiley.

Cresswell, M. J. (2000). The role of public examinations in defining and monitoring standards. In H. Goldstein & A. Heath (Eds.), *Educational standards*. Oxford: Oxford University Press.

Cresswell, M. J. (2003). *Heaps, prototypes and ethics: The consequences of using judgements of student performance to set examination standards in a time of change*. London: University of London Institute of Education.

Yaw joined CERP in 2014. He holds a PhD in Statistics, an MSc in Mathematical Sciences and a BSc (Hons) in Mathematics. His current areas of research include marking reliability, application of the Bayesian method to standard setting and test equating, and construct validity of assessment designs. Yaw has experience of analysing and modelling data in a variety of fields, and is skilled in the research and application of statistical methods. He has also taught statistics and mathematics to undergraduate students.

Like with like

Ben Jones explains why and how statistical evidence has helped stabilise standards and ensure comparability between awarding organisations and years

THE 'STANDARDS' OF A qualification are multi-dimensional. They comprise the demands of the specification (the content standard), the difficulty of a particular assessment and its associated mark scheme (the assessment standard), and students' performance at a particular grade (the performance standard) (Ofqual, 2014).

Even if there have been no changes to the content of the exam – also known as 'syllabus' and referred to here as the 'specification' – it is not straightforward to take account of the inevitable slight year-on-year variation in either the assessment standard, or the performance standard to ensure comparability. The process is especially challenging when there has been a substantial change in content and/or the nature of the assessments.

In particular, immediately after a specification has been revised, students' performances tend to decline temporarily as they, and their teachers, face some unfamiliar subject matter and have fewer support materials, such as past papers and mark schemes. Ofqual has recently noted that this phenomenon, known as the 'saw tooth effect', tends not to correct itself for at least two examination series (Ofqual, 2016). This raises an issue of fairness: it would not be fair on students who sit the first examinations of a revised specification to be awarded lower grades than those who, by accident of birth, are examined on an established specification.

Similarly, it would not be fair on students who sit a particularly demanding paper to be awarded lower grades than those who are examined on a less demanding paper. Although senior examiners aim to

produce question papers of equivalent demand between series, there are inevitably slight variations. What examiners perceive to be very minor differences in demand can, for students, be more significant, and vice versa. This is one reason why, while examiners can generally identify the existence of differences in question paper demand, they tend to be less skilled in identifying the magnitude of those differences (Good & Cresswell, 1988).

for the current entry cohort have been predicted based on the national outcomes in a reference year and taking account of any differences in the prior attainment of the current cohort (for the specification in question) compared to the reference year outcomes. (In England, 'prior attainment' is measured by mean GCSE score for AS and A-level, and by mean Key Stage 2 – exams undertaken by pupils aged 7 to 11 – score for GCSE.) Initially, these predicted outcomes were merely used as a guide, and examiner judgement was used to

The application of the comparable outcomes approach to standard setting has proved very successful in reducing grade inflation and – according to its own definition at least – demonstrably maintaining standards between years and awarding organisations

To accurately maintain the standards of a qualification by judgement alone is extremely challenging, if not impossible. Progressively, therefore, more emphasis has been placed on the use of statistical data to guide the setting of grade boundaries. Initially, this simply took the form of comparing the grade profiles of successive entry cohorts, but increasingly more refined comparisons have been made.

Starting with the 'Curriculum 2000' AS and A-levels in 2001, grade outcomes

set the grade boundaries. From 2011, outcomes were expected to be within a tolerance of the predictions, and there had to be strong justification for setting boundaries that would produce outcomes that exceeded them.

Thus, in recent years, rather than standards being defined in terms of the comparable performance of students in previous years, they have become defined in terms of comparable outcomes of students with equivalent prior attainment.

The application of the comparable outcomes approach to standard setting has proved very successful in reducing grade inflation and – according to its own definition at least – demonstrably maintaining standards between years and awarding organisations.

However, some of the assumptions underpinning the approach are quite strong. For example, it is implicitly assumed that the relationship between prior attainment and current achievement (the 'value-added rate') is consistent between subgroups. To the degree that this is not the case, a sizeable shift in the balance between subgroups in the entry cohort between years could tend the outcome towards severity or leniency.

Moreover, the comparable outcomes approach has altered the definition of grade standards and their maintenance; they are now defined in terms of outcomes, not performance. In the context of school accountability measures, this is a significant issue for some stakeholders.

These considerations have led to a rethink of how best to retain the benefits that the comparable outcomes approach has yielded, while appropriately using examiners' expertise to recognise genuine changes in performance.

As a consequence, research is being undertaken to investigate various alternatives, including how the proposed National Reference Test might best be utilised, and whether comparative judgement of students' scripts between series is a feasible way to accurately maintain performance standards (see pp. 18–19).

Defining and maintaining examination standards is a difficult task. The comparable outcomes approach has been beneficial and is likely to be used for the foreseeable future, but the current research into how it may be improved is timely and to be welcomed. ▣

REFERENCES

Good, F. J., & Cresswell, M. J. (1988). *Grading the GCSE*. London: Secondary Examinations Council.

Ofqual. (2014). *Consultation on setting the grade standards of new GCSEs in England* (Ofqual/14/5401). Coventry: Ofqual.

Ofqual. (2016). *An investigation into the 'Sawtooth Effect' in GCSE and AS/A level assessments* (Ofqual/16/6098). Retrieved from https://www.gov.uk/government/publications/investigation-into-the-saw-tooth-effect-in-gcses-as-and-a-levels

Global view

A major collaborative project, 'Setting and maintaining standards in national examinations', aims to unpack how measures and meanings differ around the world. Group member **Lena Gray** outlines some of the objectives and outcomes

THE TERM 'STANDARDS' crops up everywhere in the world of assessment. In England, awarding organisations that offer general qualifications must ensure that grades have the same meaning across subjects, in different years, and even between competing exam boards (see pp. 46–47). This is an area in which assessment researchers can see their work having real impact, and there are plenty of exciting developments to shape new thinking, as described elsewhere in this edition.

Standard setting is a topic of interest throughout the global assessment community, yet opportunities for information sharing are rare, given the politically sensitive nature of the subject. I joined the system in England in 2014, having spent most of my working life in another UK country that has a separate qualifications system, with superficially contrasting standard-setting policies and approaches. It quickly became clear to me that the opportunity to learn from other systems could be beneficial to practitioners like myself. Consequently, I am very proud to be part of a major collaborative project – 'Setting and maintaining standards in national examinations' – that aims to open conversations between international experts. Together, we are exploring how different jurisdictions tackle standard setting and maintenance for their respective national, school-leaving or university entrance, curriculum-related exams.

The project is led by a partnership of key organisations in England. Together with my Centre for Education Research and Practice (CERP) colleague Kate Kelly, I am joined by Professor Jo-Anne Baird (Chair of AQA's Research Committee and Director of Oxford University's

The project critically examines policy positions and processes for examination standards in a range of countries, drawing on analyses from in-country experts and researchers, using our fellow senior exam board personnel from around the world as participant observers

Department of Education), Dr Tina Isaacs (Institute of Education, University College London), Dennis Opposs (Standards Chair, Ofqual) and Kristine Gorgen, Research Assistant at the Oxford University Centre for Educational Assessment (OUCEA). The team brings together experience in different exam boards, qualifications regulators and academic research organisations; this provides a breadth of view that offers new insights into the field.

The project critically examines policy positions and processes for examination standards in a range of countries, drawing on analyses from in-country experts and researchers, using our fellow senior exam board personnel from around the world as participant observers. (An additional project was born out of the work, after it emerged that most of us enjoyed the privileges and constraints of insider researcher status, and that advice for our participants would be helpful. We sought and won ESRC funding to develop guidelines that have been published as 'Overcoming political and organisational barriers to international practitioner collaboration on national examination research: Guidelines for insider researchers working in exam boards and other public organisations', Oxford University Centre for Educational Assessment Report OUCEA/17/2.)

As part of the investigation, colleagues from around the world are documenting how standards are defined, and how those definitions are enacted in terms of processes and evidence used. Each system, naturally, has its own issues, and the researchers are capturing the variety of challenges they face and responses to these within their own political and economic systems. The countries involved are: Chile, England, Hong Kong, Ireland, South Korea, Sweden, France, Australia (Queensland and Victoria), US, Georgia and South Africa.

In the age of globalisation, when governments and assessment bodies around the world look to each other to question or validate their own practice, it is helpful to gain a deeper understanding of what examination standards mean in different political, social and economic contexts. The initial major outcome of 'Setting and maintaining standards in national examinations' was a three-day symposium at Oxford University's Brasenose College in 2017. This was one of the first occasions that experts had gathered together to share knowledge about the theories, policies and practices of standard setting and maintaining in their own senior school qualifications systems.

The development of this knowledge community is a critical outcome of the project, and one that myself and the other project leads have found to be enormously rewarding. There are several more tangible outcomes, too; as well as the guidelines for insider researchers mentioned above, a book – *Exam Standards: How Measures & Meanings Differ Around the World* – is in progress. We look forward to its launch during the autumn conference season. Next, we hope to work with this newfound knowledge community to produce a special issue of the journal *Assessment in Education: Principles, Policy & Practice*.

Throughout this work, it has become clear that despite differences between our systems, most awarding organisations face similar pressures and challenges. Senior school examinations shape students' future life chances, and the deeper we collaborate, the more we appreciate how vital it is to share our knowledge on how we set and maintain standards in those examinations. It is a privilege to be part of this important project. ▣

Dr Lena Gray is Director of Research at AQA's Centre for Education Research and Practice, and is an Honorary Norham Fellow of the Department of Education, University of Oxford. 'Standard setting and maintaining in national examinations' is a joint venture between the University of Oxford, AQA, IOE, and Ofqual. Lena joined CERP as Head of Research in July 2014, after many years at the Scottish Qualifications Authority (SQA).

REFERENCES

Baird, J., & Gray, L. (2016). The meaning of curriculum-related examination standards in Scotland and England: a home–international comparison. *Oxford Review of Education*, *42*, 266–284.

Baird, J., Isaacs, T., Opposs, D., & Gray, L. (Eds.). (in press). *Exam standards: How measures and meanings differ around the world*. London: IOE Press.

Gray, L. (2017). *Overcoming political and organisational barriers to international practitioner collaboration on national examination research: Guidelines for insider researchers working in exam boards and other public organisations* (Report OUCEA/17/2.) Oxford: Oxford University Centre for Educational Assessment.

Jo-Anne Baird presents initial findings at a three-day symposium at Oxford University's Brasenose College

From the archives

In each issue of *Inside assessment*, we present a previously unpublished paper on a topic of current relevance. As part of this special focus on standards, we revisit research that was shared at the 13th international conference of the International Association for Educational Assessment, held in Bangkok, 9-13 November 1987: 'Setting comparable standards on examination papers of differing difficulty', by M. J. Cresswell and F. J. Good.

Introduction by *Ben Jones*

PUBLIC EXAMS IN THE UK have recently undergone a period of reform, instigated by government changes to the system. CERP research has helped inform the new generation of GCSE, AS and A-levels developed by its parent organisation AQA. GCSEs are returning to a linear structure and will adopt a nine-point grade scale. GCEs are also becoming linear, with the AS qualification being decoupled from A-level.

Reform to qualifications on this scale is not unprecedented: pilot joint examinations of the 1970s brought together CSE and O-level standards. Subsequently, GCSE grade criteria for nine subjects were developed and passed on in 1986 for use by the awarding bodies. Assessment research at this time sought to explore differentiation by task: the relatively new idea that exam papers should target different levels of achievement. Each candidate chose to take a particular combination of papers, and those taking harder ones were eligible for higher grades; some grades were available from more than one combination of papers. This meant that comparable grading standards had to be set on papers of differing difficulty – a framework that awarding organisations continue to work within (see pp. 26–28).

Mike Cresswell and Frances Good investigated the ability of suitably qualified judges to set standards that take differing difficulty of papers into consideration. It had been expected that the judges would make decisions that allowed for differences in the difficulty of the papers so that candidates would have, on average, an equal chance of meeting any given grading standard, whichever papers they took. In fact, the judges tended to set standards so that, when

candidates were entered for both an easy and a hard paper, more of them were deemed to meet the prescribed standard on the easy paper. This phenomenon became known as the Good & Cresswell effect, which is still referred to today.

Cresswell and Good's work in this area – shared below – provided the context for a lively discussion about the processes used by standard setters to reach fair decisions. It set the tone for a range of subsequent research projects that opened up understanding about the use of judgement in awarding. Thanks to this work, and that of researchers thereafter, improvements to the standard-setting procedures have been made. ▣

Setting comparable standards on examination papers of differing difficulty

By M. J. Cresswell and F. J. Good

Presented at the 13th international conference of the International Association for Educational Assessment, Bangkok, 9-13 November 1987, on behalf of the Associated Examining Board (AEB, a predecessor body of AQA). Abridged for *Inside Assessment*, 2018

Introduction

The CSE and GCE O-level examinations taken by pupils at the end of compulsory schooling in England and Wales are to be replaced from 1988 with new GCSE examinations [see pp. 46–47 for contextual information about this qualification]. One of the distinctive features of GCSE in some subjects is that candidates will have to choose to enter for a particular combination of papers. Different combinations of papers will be set at different levels of difficulty and give access to different, but overlapping, ranges of grades. Therefore, it is necessary to set comparable grading standards on examination papers of differing difficulty. This paper reports some work that investigated the ability of suitably qualified judges to manage this task.

Background

One of the ways in which GCSE papers of different difficulty – referred to as 'differentiated' papers – are organised is by using four papers of which each candidate takes two: paper 1 and paper 2, paper 2 and paper 3, or paper 3 and paper 4. The papers are all designed to be of different difficulty: paper 1 being the easiest and paper 4 the hardest.

The GCSE grade scale extends from A-G; candidates who fail to achieve a G are treated as ungraded and do not receive a certificate. Candidates taking papers 1 and 2 are eligible for grades in the range E-G, those taking papers 2 and 3 are eligible for grades in the range C-F, and those taking papers 3 and 4 are eligible

for grades A-D. This particular arrangement of papers is the one used in the work featured here. There are two ways in which an examination of this type might be graded:

(i) the mark scales of the papers of different difficulty might be equated so that all candidates are positioned on a common aggregate mark scale that is then partitioned, using professional judgement, into grades;

(ii) the three different combinations of papers are graded separately, relying upon the ability of those fixing grade standards to set comparable standards for grades available via more than one combination.

At the time of writing (Autumn 1987), most, if not all, examining groups ► that are to award GCSEs favour the second approach. This preference is based partly on an appreciation of the theoretical problems of equating in this context (Cresswell, 1982) and partly on a desire to use straightforward procedures that do not involve adjusting candidates' marks. Therefore, it is of pressing importance to ensure that suitably qualified judges can set comparable standards on papers of differing difficulty.

Method

Experimental examinations in History and Physics were devised. These each involved four papers (as described above) and were taken by candidates shortly before they sat their normal examinations. The number of candidates taking each paper is shown below:

TABLE 1

No. of candidates taking each paper

	Paper 1	Paper 2	Paper 3	Paper 4
History	157	105	137	53
Physics	132	149	204	71

The completed examination scripts were marked and graded. For each paper, the lowest mark that could be considered to qualify for certain defined grades was determined by professionally qualified judges. In each subject, eight judges were organised as two teams of four; these teams graded the papers independently.

The judges were people who had previously been involved in public exams and had experience of setting grade standards. They were not given explicit criteria to use; instead, their tacit knowledge (Sadler, 1987) and intuition, which are assumed to reflect some general educational consensus, were the basis for their judgements. This approach (see Christie & Forrest, 1981) is the one conventionally used to set standards in British examinations at present (1987), although attempts are now being made to move towards the use of more explicit criteria, notably in Scotland (Long, 1985).

TABLE 2

Minimum marks in grades – History

Paper	Grade	Min mark* in grade: Team 1	Min mark* in grade: Team 2	Difference (T1-T2)
1	F	21	22	-1
2	F	23	22	+1
	C	35	34	+1
3	F	18	23	-5
	C	29	34	-5
	A	41	40	+1
4	C	23	21	+2
	A	32	35	-3

* Max. mark for each paper = 50

TABLE 3

Minimum marks in grades – Physics

Paper	Grade	Min mark* in grade: Team 1	Min mark* in grade: Team 2	Difference (T1-T2)
1	F	38	41	-3
2	F	28	31	-3
	C	64	64	0
3	F	36	29	+7
	C	59	60	-1
	A	76	79	-3
4	C	46	46	0
	A	64	57	+7

* Max. mark for Papers 1-3 = 100
Max. mark for Paper 4 = 85

Results

Tables 2 and 3 show the decisions of the teams of standard setters for each paper.

For standard-setting purposes, the level of agreement found between the two teams of judges in each subject seems tolerable. Few of the discrepancies between the teams are large compared with the disagreements found between different markers of the same script. Because the two teams of judges in each subject reached a reasonable agreement, their results were pooled to give a single set of minimum marks for grades. Table 4 shows the effect of these pooled results in terms of the candidates' performances.

For all except one of the six pairs of papers, the judges' decisions enable more candidates to meet a given standard on the easier paper than are able to meet what is nominally the same standard on the harder paper. This effect has also been observed elsewhere in examinations that use differentiated

TABLE 4A

Percentage of candidates meeting grade standards set on each paper in History

Papers	Grade	1 (Easy)	2	3	4 (Hard)
1 & 2	F	47.2	38.0		
	C	*	3.5		
2 & 3	F		54.7	46.2	
	C		7.6	5.7	
3 & 4	C			42.9	23.0
	A			10.6	2.8

TABLE 4B

Percentage of candidates meeting grade standards set on each paper in Physics

Papers	Grade	1 (Easy)	2	3	4 (Hard)
1 & 2	F	72.2	86.7		
	C	*	1.1		
2 & 3	F		97.9	74.2	
	C		11.3	3.1	
3 & 4	C			46.2	19.1
	A			10.4	2.9

* standard not set.

papers (Good & Cresswell, in press). One purpose of the standard-setting process is to compensate for the differences in difficulty between the papers; it is to be expected, for instance, that a given standard will represent a higher proportion of marks on an easy paper than on a hard one.

With examinations that use differentiated papers, if the standards set do not make appropriate allowance for the differing difficulties of the papers, the consequence is that where a grade is available from more than one combination of papers, it will be more difficult to obtain that grade via one route than via another. In the present study, some candidates completed three papers so that it was possible to compare the grade they achieved from two different versions of the examination. Because of the relatively small numbers of candidates involved, and the similarity of the data in Physics and History, data from the two subjects have been combined. These data are reported in Table 5. (Details of the processes by which the grades were determined, based upon the judgements reported in Tables 2 and 3 are given by Good & Cresswell, in press).

The data show that when the same candidates were examined on two different combinations of papers, there was a tendency for them to receive higher grades from the easier combination. This is the direct result of the tendency, evident in Table 4, for the standard setters to be more lenient on the easier papers.

TABLE 5 Comparison of grades awarded via different combinations of papers

Type of result	Number of candidates
same grade from both versions taken	66
higher grade via harder version	6
higher grade via easier version	27

Discussion

Results reported in this paper imply that the teams of standard setters applied a performance-based notion of standards to this task, i.e. fixing the lowest mark in each grade by envisaging candidates who just qualified for that grade in previous examinations and attempting to judge what score they would get on each paper – as used in general standard-setting purposes (Livingston & Zieky, 1982).

The judges tended, initially at least, to scrutinise scripts in a partial fashion. For each grade standard fixed, they paid particular attention to 'relevant' questions, i.e. those that they felt would discriminate between the grade in question and the ones below. For grade C, for example, they looked at questions that they expected a minimally qualified grade C candidate to tackle with some, but not complete, success.

On the easier papers, the judges argued that only some of the questions were relevant in this way to any particular grade. Nonetheless, they generally accepted that the candidates' grades should be determined by their total scores and that the minimum mark for each grade should reflect performance on all the questions in the paper.

Where a grade is near the top of the range of achievement covered by a paper, it is likely that the other questions in the paper will be easier than the 'relevant' questions; where the grade is one of the lowest awarded on the paper, the other questions will tend to be more difficult than the 'relevant' ones. In the present study, the judges did seem to be constrained by the nominal equality of the marks; for a given grade, they appeared unwilling to require a much higher proportion of marks on the easier set of questions than on the harder set. However, candidates were, in general, gaining their marks in an unbalanced way and obtaining a considerably greater proportion of easier marks than harder marks. Thus, the judges may have underestimated the total number of marks likely to be scored by candidates getting grades near the top of the range and, conversely, to have overestimated the marks likely to be scored by candidates getting grades near the bottom of the range covered by the paper.

Although this discussion is based on observation of the judges, there is an element of speculation. The assumption that the judges effectively overvalued the candidates' achievements on the easier papers and/or undervalued them on the harder papers can be challenged. A second explanation is based on the assumption that, for any particular grade, different proportions of candidates are genuinely able to demonstrate the required level of achievement on the different papers; the judges are not, in this view, wrong. For an

On the easier papers, the judges argued that only some of the questions were relevant in this way to any particular grade. Nonetheless, they generally accepted that the candidates' grades should be determined by their total scores and that the minimum mark for each grade should reflect performance on all the questions in the paper

explanation of this type to hold, the judges would have to adopt a more criterion-based approach. Rather than attempting to estimate the overall performance of a minimally qualified candidate, the judges would have in mind a set of criteria, which, if met by a candidate's work, imply that the work as a whole meets a given standard. The overall standard is defined as meeting all the criteria, but some of the criteria may, in fact, be met by candidates whose work as a whole does not reach the standard required.

In differentiated examinations, the differences in difficulty between the papers arise because different weights are given to the various assessment objectives specified for the subject as a whole. Therefore, from this perspective, it is not unreasonable for different proportions of candidates to meet the criteria for a given standard on different components of a differentiated examination. Further, it is probable that the easier papers of a differentiated papers examination will test those aspects for which a relatively large proportion of candidates can meet the criteria. The problem with this is that the criteria assessed by each version of the examination are only a subset of the full set of criteria for the standard concerned in the subject overall.

Candidates taking an easier combination of papers (e.g. papers 1 and 2) are not measured against the criteria that fewer candidates meet, and candidates taking a harder combination (e.g. papers 3 and 4) are unable to demonstrate their achievement in terms of the criteria that more candidates meet.

Thus, although the criteria may be appropriate for the award of the particular grade concerned, the subsets of those criteria that are actually brought to bear upon candidates' achievements within the different combinations of papers that comprise the examination do not make

comparable demands. As a result, the standard that is set on one combination of papers will not be comparable to what is nominally the same standard set on a different combination.

When suitably qualified judges set a standard on an examination paper, they are given the task of recognising candidates' work that merits the award of the grade in question. With a criterion-based approach, each candidate's work is judged against the criteria defining the grade. However, the candidates' work can be viewed in different ways, perhaps giving different decisions as to the standard it reaches. The judges might adopt a strategy that can be characterised as the *script as artefact* or they might consider the script as response.

The strategy of script as artefact involves the judges scrutinising candidates' scripts to determine the presence or absence of particular qualities. The criteria used with this strategy refer only to the characteristics of the scripts themselves. Criteria of this type make no reference to the tasks that the candidates were asked to perform: that is, to the questions in the examination paper. There is an implicit assumption that the nature of the questions does not affect the meeting of the criteria: that opinion is equally easy to recognise in any text; that ideas of ratio are not more easily applied in some contexts than in others; and so on.

The strategy of *script as response* offers a theoretical solution to this problem. With this strategy, standard setters attempt to judge the candidates' scripts in the context of the particular questions set; they attempt to assign a grade representing not the quality of the script per se, but the quality of the script considered as a response to the particular examination paper concerned. However, it is much more difficult to formulate standard-setting criteria for use under the script as response strategy since such criteria must, of necessity, refer to the characteristics of the questions as well as the scripts.

Furthermore, it is notoriously difficult to predict, in any particular instance, precisely how the characteristics of a question influence the ease with which a given skill or item of knowledge can be demonstrated; this is, of course, the practical problem that undermines the script as artefact strategy. Although these issues are currently being actively investigated (Pollitt et al, 1985), it is difficult to see how, given the present understanding of the relationships involved, criteria for use under the script as response strategy could be formulated. The question that then arises is whether the strategy can operate in the absence of explicit criteria.

Conventionally, it is argued that standard setters can gain extra insight

into the ways in which questions have functioned by perusing candidates' scripts. This is no doubt true, but the circularity of the argument is clear. If, in a particular paper, fewer candidates than usual demonstrate a particular skill, is this an indication that the context used to test that skill imposed unusual demands, or does it mean that few of the candidates concerned possess the skill in question? It is impossible to decide, but under the script as response strategy, the value of candidates' responses in terms of grades depends crucially upon the answer to this question.

It is worth noting that procedures vary considerably between examining authorities, and from subject to subject, making categorical statements about them unwise. Nonetheless, it is probably the case that most British examination boards' traditional procedures imply a preference for the strategy of script as response [at the time of writing]. However, exceptions to this do occur; for example, when it is thought that the detail of the particular examination has little influence upon the quality of candidates' responses. For example, in Art, standard setting is usually a matter of comparing each candidate's work with specimens of work from previous examinations at each grade boundary.

In examinations involving differentiated papers, where the difficulties of the papers vary, the choice of standard-setting approach is critical. It has already been pointed out that all the criteria that comprise the standard for a grade must be assessed by any combination of papers that permits the award of that grade. If this requirement is met for grades that are available on more than one combination of papers then, for those grades, the more difficult combinations of papers cannot be more difficult because they cover more difficult, but relevant, criteria: they must be more difficult because of the particular contexts in which the relevent criteria are assessed. In these circumstances, a script as response standard-setting strategy ensures comparability of standards across the different papers and, subsequently, across the different versions of the examination.

Conclusions

There is a tendency for judges setting nominally the same standard on papers of differing difficulty to make decisions that enable more candidates to reach the standard on easier papers. Two mechanisms to explain this phenomenon have been suggested. One shows how judges might make biased decisions when attempting to predict the likely scores of minimally qualified candidates. The other mechanism assumes that the judges operate a set of performance criteria that must be met for

the award of any given grade. It has been pointed out that even the correct use of such criteria need not lead to comparable standards being set on papers of differing difficulty. Even if all the criteria that comprise the standard for a grade are assessed by each combination of papers that permits the award of that grade, comparable results cannot be guaranteed. This is because the contexts in which the criteria are assessed will tend to be more difficult in the harder papers.

Standard setters working with papers of differing difficulty must therefore interpret any performance criteria they use in a such a way as to make allowance for this effect. In essence, this requires them to adopt a performance-based approach (in which, for each paper, they attempt to predict the likely score of minimally qualified candidates) rather than a fully criterion-based approach (in which candidates scripts per se are judged in the light of a set of explicit or implicit performance criteria).

In conclusion, it seems clear that until the processes involved are better understood, fundamental improvements in standard-setting procedures will not be possible. Intuition and tacit knowledge are likely to remain the principal tools used. In these circumstances, the ability of suitably qualified judges to carry out the task required of them in the present study – to set comparable standards on examination papers of differing difficulty – must be in doubt. ■

REFERENCES

Christie, T., & Forrest, G. M. (1981). *Defining public examination standards.* Basingstoke: Macmillan.

Cresswell, M. J. (1982). Some possible approaches to the problem of examining across a wide range of ability. *Curriculum*, *3*(2), 38–44.

DES. (1985). *GCSE national criteria for mathematics*. London: Her Majesty's Stationery Office.

Good, F. J., Cresswell, M. J., & Sudway, H. (1986). *Novel examinations at 16+ Research Project: Pilot Study Report.* London: Secondary Examinations Council

Good, F. J., & Cresswell, M. J. (in press). Grade awarding judgements in differentiated examinations. *British Educational Research Journal*.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton: Educational Testing Service.

Long, H. A. (1985). Experience of the Scottish Examinations Board in developing a grade-related criteria system of awards. Paper presented at the 11th international conference of the International Association for Educational Assessment.

Murphy, R. J. L. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology*, 52, 58–63.

Pollitt, A., Hutchinson, C., Entwistle, N., & De Luca, C. (1985). *What makes exam questions difficult?* Edinburgh: Scottish Academic Press.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, 13, 191–209.

SEC. (1985a). *Report of the Grade-related Criteria Working Party for English.* London: Secondary Examinations Council.

SEC. (1985b). *Report of the Grade-related Criteria Working Party for Mathematics.* London: Secondary Examinations Council.

Contextual notes

IN ENGLAND, STUDENTS generally receive their first academic qualification, the General Certificate of Secondary Education (GCSE), at the age of 16. While other countries reward completion of secondary education via a diploma that summarises achievement across the curriculum, a GCSE is awarded for each individual subject. A student will take GCSEs in a number of subjects, some of which are compulsory – for example, English, Mathematics and Science. These are taken alongside optional subjects, including History, modern foreign languages, and subjects such as Food Preparation and Nutrition.

After receiving their GCSEs, many pupils decide to continue their studies and select three or four subjects to study for another two years. If successful, they obtain further subject-specific qualifications, called (GCE) A-levels, which can be used to apply to university.

Maintaining standards means ensuring that students displaying similar levels of ability in a subject receive the same final grade, even if examinations differ across subjects and over time. In England, standards are maintained using two sources of evidence: experienced examiners' judgement and statistical data. The process is known as 'awarding'. Statistical evidence relies on cohort comparability: it is assumed that the average relationship between student prior and final achievement does not change if the cohort composition remains the same. To set A-level standards, for instance, the relationship between GCSE and A-level results of a previous reference cohort is applied to the following years. This relationship is expressed by the mean GCSE score of the reference group, split by decile, crossed with their A-level results. The resulting percentages are then used to set future A-level cut-scores, by ensuring that successive GCSE to A-level relationships match the reference one as closely as possible.

To account for fluctuations in cohort ability, the comparable outcomes approach uses 'predictions' (based on

candidates' prior attainment) to determine how well we might expect this year's cohort to do. These predictions consider the relative ability of last year's cohort and this year's cohort, as well as the performance of last year's cohort. Essentially, we predict the percentage of candidates we expect to get a given grade this year by comparing their ability with the ability of last year's candidates – and how well that prior cohort did on last year's exam.

We use this predicted grade distribution for the current cohort to work out statistically recommended boundaries (SRBs) – the positions where grade boundaries should be placed to produce outcomes that are comparable to those the prior cohort achieved in last year's exam. But we don't set grade boundaries purely based on these statistical predictions. In awarding meetings, a committee of subject experts meets to agree the grade boundaries for a given specification. While the committee members use the SRBs as a starting point, they can deviate – to a degree – from the statistics. In fact, a key aspect of their role is to account for any changes in the standard that might not have been captured by the statistics.

The resulting evidence – statistical and judgemental – is discrete. The assessment researcher's challenge stems from understanding how the statistical information (i.e. comparable outcome prediction) and the professional judgements of candidates' work can be combined to set grade boundaries. In practice, we tend to favour one of the sources of evidence at the expense of the other. This is largely dependent upon how robust we know the statistics are (i.e. how many candidates the predictions used to derive them are based upon).

CERP researchers seek to build on current awarding procedures to meet the challenge of setting grade boundaries that most clearly reflect prior information on the anticipated cumulative percentage outcomes (i.e. grade distribution) at subject level, as well as appropriately incorporating professional judgements. Past and present work undertaken in this field is featured throughout this publication. ■

Meet the researchers

The **Centre for Education Research and Practice (CERP)** comprises a range of specialists, from statisticians and psychologists, to educationalists and scientists. In the first of our regular series of researcher profiles, we introduce you to the team behind the expertise. In this 'Standards' issue, we meet the staff involved in the awarding process

Ben Jones
Head of Standards

Ben first joined the Joint Matriculation Board (JMB), a predecessor of AQA, in 1990 and held a variety of posts as the organisation merged and developed. After a two-year career break in 2005, he returned to his current post, the primary responsibilities of which are to ensure the effective provision of AQA's awarding process and the integrity of the standards of its qualifications. Ben represents AQA on the Joint Council for Qualifications' (JCQ) Standards and Technical Advisory Group, and on Ofqual's Standards and Technical Issues Group.

Before joining the organisation, Ben was Adviser in Educational Assessment to the government of Tonga for three years and spent five years as a Research Associate in the Division of Education at Sheffield University. There, he worked on various research projects including the initial investigation into DES Performance Tables for schools (the Contexts Project) and the first years of the National Youth Cohort Study. Ben has a BA in Economics and an MSc in Social Research Methods.

What's the best part of working in standards?
Working on a varied and intellectually demanding enterprise with a team of great colleagues.

Describe a notable research highlight.
The Contexts Project (see above). The 1980 Education Act required schools to publish their examination results for the first time. Thereafter, unofficial league tables were compiled by newspapers. We were the first researchers to contextualise schools' results by controlling for their students' prior ability.

Years spent in standards: 26

Simon Eason
Principal Research Manager

Simon joined the Associated Examining Board (AEB), a predecessor of AQA, in 1987. His work involves the use of prior attainment data in calculating overall expected subject outcomes. Simon serves on the AQA Standards Unit, which advises on the maintenance of awarding standards, and he is responsible for AQA's published results statistics and internal statistical archives. He undertakes regular statistical analysis in support of the JCQ Standards and Technical Advisory Group. Simon obtained his BSc in Mathematics, Statistics and Computing from Thames Polytechnic and his MSc in Management Science and Operational Research from Warwick University.

Tell us about the AQA Standards Unit.
The standards unit was originally set up when the constituent boards of AQA (JMB/NEAB and AEB/SEG) merged in the late 1990s to form AQA; its purpose was to ensure consistency of standards. Although the work has evolved since then, the main focus remains to ensure that the standards of AQA's qualifications are maintained over time.

Which work has meant the most to you?
I have enjoyed developing the statistical models that use prior attainment to calculate overall expected subject outcomes. The general methodology was first introduced by Mike Cresswell for the Curriculum 2000 GCEs when they were awarded for the first time in 2001 (AS) and 2002 (A-level). Having been involved in the initial implementation, I developed the methodology to be used for GCSE qualifications, which has been adopted by the wider awarding community.

Years spent in standards**: 31**

Lesley Meyer
Senior Researcher

Lesley joined the AEB in 1998, and is heavily involved in the maintenance and development of AQA's awarding procedures. Lesley worked for eight years as a medical statistician in various areas of medical research, including public health, diabetes, and breast and prostate cancer. Throughout this period, she also taught medical students, both at Undergraduate and Master's level, which stimulated her interest and eventual move into education.

Lesley completed her BSc degree in Pure Mathematics and Statistics at Royal Holloway and Bedford New College (London University) and her MSc in Medical Statistics at Southampton University.

What's the most interesting piece of standard-setting work you've undertaken?
A paper that looked into which grades should be judgemental at A-level and GCSE. These grades are different for each examination type and depend on various factors, including: importance in terms of selection for and progression to a higher level of study or the workplace; the likelihood of producing the most reliable judgements; and the method of aggregation used to combine unit (or component) marks to subject level. My paper reviewed why the current judgemental grades are used in A-levels and GCSEs, in particular, and whether they should be altered in the new specifications.

What is a key feature of your role in awarding?
As Awarding Coordinator, I am in continuous liaison with the various teams in the Operations department at AQA. During each awarding series, I am in (what seems like) constant conversation with colleagues in preparation and marking teams to catch up with the data and organisational situations on each award. Also, throughout the year, I work closely with the members of Planning and Resource Management who put together the marking schedules, standardisation and awarding documents for every specification. These opportunities to liaise closely with other teams broaden the scope of my role and are very rewarding.

Years spent in standards: In November this year it will be 20.

Martin Taylor
Senior Researcher

Martin's background differs from that of other members of AQA's Centre for Education Research and Practice (CERP) in that his previous experience was largely in teaching rather than research. Most of his teaching experience was in two sixth-form colleges and included the role of Head of Mathematics. He joined the Research team of the then Associated Examining Board (AEB) in 1991.

Martin's work tends to focus on technical issues associated with current and proposed examinations, with particular reference to internal assessment.

What is the most rewarding part of working in standards?
Collaborating with Ofqual and other awarding bodies in the interests of maintaining comparable standards at national level.

What are the key features of your work on internal assessment?
Internal assessments, which are marked by teachers, have to be moderated by the awarding body to check that they are in line with national standards. I ensure that the process is statistically robust while being manageable.

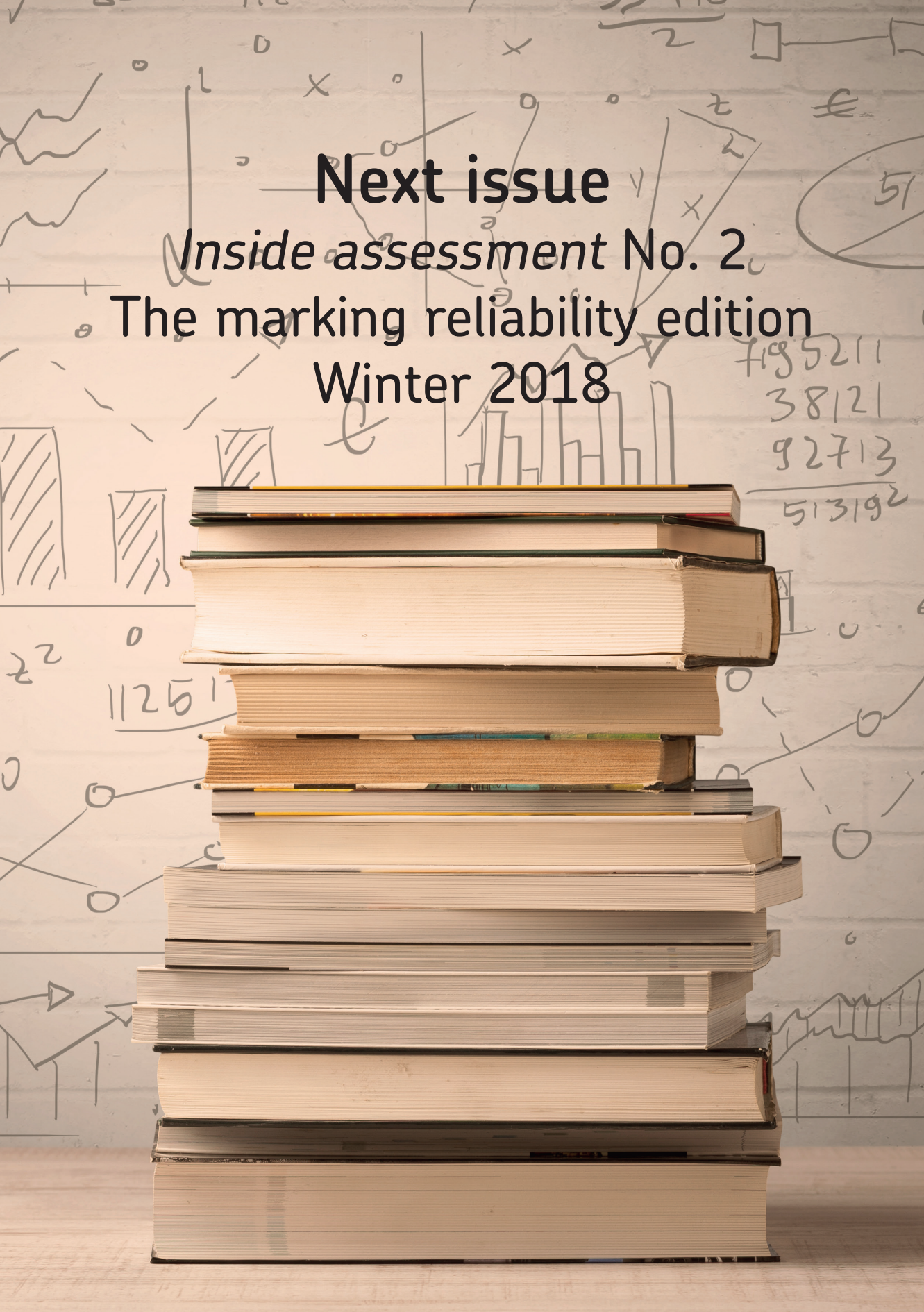
Years spent in standards: 27

Next issue

Inside assessment No. 2

The marking reliability edition

Winter 2018



Centre for
Education Research
and Practice (CERP)

AQA
Realising potential