RESPONSE TO THE EDUCATION SELECT COMMITTEE INQUIRY INTO THE ADMINISTRATION OF EXAMINATIONS FOR 15-19 YEAR OLDS IN ENGLAND

FROM

CENTRE FOR EDUCATION RESEARCH AND POLICY (CERP)

7 November 2011

CERP has a proven record of high quality research in education assessment, undertaken by it and its predecessors for over forty years. Legally, CERP is part of AQA, but retains its own identity and research freedoms.

We use an appropriate range of quantitative and qualitative methods to maintain the highest standards of academic rigour whilst being grounded in the practical realities of assessment; this makes our recommendations evidence-based, relevant and manageable. Our work is meticulously reviewed by a prestigious committee of national and international experts, chaired by Professor Jannette Elwood of Queen's University, Belfast. Our work is disseminated at national and international conferences, published in academic books and journals, in reports for policy-makers, and through newsletters and magazine articles for practitioners.

1. INTRODUCTION

1.1. This response, which relates solely to general qualifications (particularly GCSE and GCE), focuses on the Select Committee's query "how to ensure accuracy in … marking scripts and awarding grades". It particularly addresses the question of whether standards are being maintained over time and between awarding bodies (ABs), and, if not, whether a different AB structure would help.

1.2. We begin with a brief but necessary preamble, to clarify key terms. We then acknowledge some of the issues identified by the Select Committee, but argue that the cause attributed is incorrect and so the solution, of restructuring awarding bodies, would not be effective; rather, many of the benefits of the system would be lost and major risks introduced.

PREAMBLE

1.3. **Marking and grading**. Students' examination scripts are marked against defined mark schemes and, whilst question papers and mark schemes are broadly consistent in demand, this inevitably fluctuates slightly between examination series. Grade boundaries are, therefore, set to compensate for such minor variations and ensure that standards are maintained. Taking advice from their Senior Examiners, subject Chairs recommend sets of grade boundaries to the AB's Responsible Officer[i], whose decision is final[ii].

1.4. **Statistics and judgement**. Two sources of evidence determine the correct grade boundaries. Statistical data have become increasingly sophisticated over the years and now play a

major role in awarding, affording greater confidence that standards are being maintained. In addition, Senior Examiners use their expert judgement to scrutinise scripts, comparing them with those of the previous year.

1.5. **Grading and content standards**. The term "standards" has several meanings in education (Baird, 2007), and is used in two ways in the field of assessment. Grading standards refer to whether a student has an equal chance of being awarded a given grade in one examination, as in another. This is essentially what ABs aim to achieve in their awarding meetings (above). In the wider context, content standards refer to the criteria of what is being assessed and are far harder to compare. It is natural that content standards will change, especially over long periods of time: the content standards of computing qualifications, for example, are higher now than ten years ago. Over shorter periods of time and between ABs whose specifications are tightly regulated, variation in content standards is minimal, and thus large changes in outcomes would not be expected.

2. THE ISSUE

2.1. The focus of the Select Committee suggests that there is a perceived problem with the current marking and grading processes, which might be solved by restructuring awarding bodies.

2.2. Each summer sees allegations that rises in GCSE and GCE grade outcomes are inflationary, and therefore unjustified. Each summer, there is an increase in the number of enquires about results[iii] (requests for re-marks), implying a reduction in marking reliability. Undoubtedly, these trends reduce confidence in the examination system. This inquiry provides a unique opportunity to review empirical research evidence regarding the quality of marking and grading in the English examination system.

3. THE CAUSE?

Marking reliability

3.1. In order to maximise marking reliability, examiners are well trained at standardisation meetings and scripts are marked to as tight a mark scheme as is appropriate to the subject, without compromising the validity of its assessment[iv]. The marking reliability of GCSEs and GCEs is comparable with that observed for similar international assessments (Meadows & Billington, 2005), even for assessments requiring relatively subjective judgements, for example GCSE English (Fowles, 2009).

3.2. It is likely that the trend of increased enquires about results reflects not a reduction in marking reliability, but an increase in the high-stakes nature of general qualifications. Indeed, evidence suggests that re-mark requests are highly targeted, involving students who have just fallen short of a higher grade, and do not decrease even when there is research evidence of improved marking reliability (Meadows and Taylor, 2008). Overall, the proportion of re-marks leading to a grade change has not varied greatly over the past five years[v].

3.3. It is difficult to see how levels of marking accuracy could be improved by changing the current AB structure. Indeed, evidence suggests that the competition that exists within the industry serves to stimulate better assessment practices. Online marking, which has recently been introduced on a large scale, requires examiners to mark by question rather than whole script. This has served to reduce bias, focus expert examiners on the more complex questions and thus substantially improved marking reliability (Fowles, 2005; Pinot de Moira, 2009; Taylor, 2007). Online standardisation has also significantly improved marking reliability by ensuring consistency in examiner training (Chamberlain & Taylor, 2010). These innovations are products of healthy

competition between ABs in the reliability and validity of assessment offered to teachers and students.

Standards maintenance

3.4. The conclusion often drawn from rising grade outcomes is that they are unjustified and result from ABs competing for entries, a further implication thus being that standards between ABs are out of alignment[vi]. The key questions are whether unjustified grade inflation exists in the system and, if it does, is it demonstrably because there are multiple ABs? Would the establishment of a different organisational structure in itself eliminate grade inflation, whilst resolving the issues associated with genuine grade increases? The evidence of the following sections suggests that the answer to these questions is "no", and that – given the rigour of the current grading system – confidence in standards in general, and in inter-year and inter-AB comparability in particular, should be high.

3.5. ABs focus primarily on maintaining standards between years and between themselves. (Inter-subject comparability, whilst important, is harder to ascertain unequivocally, both judgementally and statistically – see Coe, 2007.) Although the awarding process is subject to external regulation, most obviously via the statutory Code of Practice (2011), ABs themselves are rigorous both in monitoring standards and introducing improved approaches to ensure their integrity[vii]. For example, CERP made a substantial contribution to the seminal Qualifications and Curriculum Authority book, which comprehensively covered the history and practical and theoretical issues surrounding setting and maintaining standards in curriculum-embedded general qualifications such as GCSE and GCE (Newton et al., 2007).

3.6. Much of CERP's work concerning maintenance of standards, particularly that of Good and Cresswell (1988), Jones (2009) and Stringer (in press) demonstrates that, without adequate statistical guidance, even experienced examiners find it hard to set grade boundaries that accurately take into account variations in question paper demand. Consequently, statistical guidance has come to play an increasingly important role in maintaining standards, especially since longitudinal student-level data became available (mapping student attainment at Key Stage 2, GCSE and A-level).

3.6.1. The outcomes of the previous year, adjusted for any changes in the ability profile of the student entry, provide a prime source of evidence for **maintaining inter-year standards**. Such measures, which are inflation-proofed, have been adopted by all ABs, have proven to be accurate, and have been instrumental in minimising grade inflation (Stacey, 2011). Ofqual recently commissioned NFER to undertake an independent evaluation of this approach; its positive report is available on the Ofqual website[viii] (Benton & Lin, 2011).

3.6.2. **Inter-AB standards** are also maintained via this approach, since the national outcome for one year forms the basis of the next year's subject outcomes. This use of a common, national basis promotes comparability, as far as ABs meet their expectations. Almost invariably they do: for example, in summer 2011, nearly all A-level outcomes at grade A were within +/-1% of their expectation. A post hoc comparability check is also undertaken, again taking into account the ability profile of the subject entries. The 2011 results are awaited, but in 2010 the outcomes of very few subjects were more than 1% from their prediction, and most of these were small-entry subjects for which statistical analyses are less reliable.

3.7. Grade boundaries are not, however, determined solely by statistics. Responsible Officers will approve awards which are not in line with statistical expectations if persuaded that the judgemental evidence of their Senior Examiners indicates that the correct standard lies elsewhere.

In these instances, the judgemental evidence (detailed descriptions of student performance compared with that of the previous year) is carefully documented.

3.8. In addition to these rigorous monitoring procedures and checks, ABs undertake bespoke research, cross-moderation and comparability exercises on any subject which gives a cause for concern[ix] and, drawing on its annually-collected national archive material, Ofqual undertakes 5-yearly reviews of standards. All of these measures are robust, effective and efficient.

3.9. Finally, ABs maintain a strict firewall around data which are collected for standards/assessment purposes: in particular, they are not shared with their marketing departments, nor used for commercial purposes. This restriction is reflected in the JCQ's strict protocol governing the use of these data[x].

3.10. Given these robust procedures, why have there been increases in grade outcomes? In recent years this can be attributed to a range of factors, not least increased investment in education in real terms (which has not, incidentally, resulted in pro rata grade increases), the increased pressure on schools to improve their performances for DfE league tables, and the consequential improvement in teaching and learning. If the government's target of all schools having 50% of students achieving 5+ A*-C grades at GCSE by 2015 (up from 35%) is to be met, then significant genuine increases in grade outcomes will have to be realised.

3.11. There are also some arguably less positive reasons for the increases, including increased teacher focus on students on the borderline of grades (Richmond & Freedman, 2009), teaching to the test (Boyle & Charles, 2011) and use of past examination papers and mark schemes to 'coach' students (Daly et al., in press). The culmination is that while performance on the highly defined set of knowledge and skills embodied by GCSE and GCE improves, performance on measures of wider learning such as those included in international surveys, may decline (e.g. Bradshaw et al, 2010).

4. THE SOLUTION

4.1. In GCSE and GCE a small degree of marking inaccuracy is an inevitable price for the delivery of a valid assessment (Meadows & Billington, 2005). There is a trade-off between reliability and validity, and these qualifications are respected for focusing on the latter, whilst aiming to maximise the former. Absolute marking reliability would only be feasible by, for example, using multiple choice tests, but this would significantly reduce assessment validity – our obligation is properly to assess the required knowledge and skills[xi].

4.2. Thus, insofar as marking accuracy is an issue, the solution lies not in changing organisational structures, but in measures such as further improving examiner standardisation, quality assurance and monitoring, activities in which all ABs are fully engaged. The on-going issues related to the marking of the KS2 National Curriculum Tests are a reminder that centralisation does not eliminate challenges to marking quality.

4.3. Similarly, increasing grade outcomes are not due to the existence of several ABs, nor to competition between them. Cresswell (1995) demonstrated from a theoretical stance that it was not in the individual or collective interests of ABs to compete on standards, a finding supported by Malacova and Bell (2006) and by the outcomes of the now routine annual analyses.

4.4. Inter-year and inter-AB standards can now be precisely set, evaluated and monitored. Increasing outcomes are a feature of the curriculum-embedded assessments, not the structure of the system. Even systems with a single AB are prone to rising grade rates and will thus face the issue of eroding discrimination in, and fitness for purpose of, their qualifications. The National Curriculum Tests are a case in point; rising outcomes are an inherent feature of maintenance of standards where a clearly defined set of knowledge and skills is being assessed.

4.5. The benefits of the current system are manifold, firstly and perhaps most importantly in the area of research and development. Innovative syllabuses and assessment schemes can flourish. "Bottom-up" developments are demand-led and implemented by groups who intimately know the curriculum requirements of, and are enthusiastic about, their subject areas. These were more common in the past[xii] but still continue[xiii].

4.6. Second, the customer – whether teacher or student – benefits from having several ABs to choose between in areas such as price, service, features of the specification and quality of the support materials. Competition in these areas, which is increasing, ultimately enhances the quality both of students' education and their assessment. It encourages ABs to adopt a progressive, "can do" mentality – not a safe, entrenched mindset which tends towards retaining the status quo. ABs are always looking to draw on their substantial technical expertise to improve assessments initially for the benefit of the teachers and students taking their examinations, but ultimately of the whole system.

4.7. Third, any assessment system – whether single or multi-organisational – is subject to risk. Current measures to counter misalignment of standards are robust, efficient and effective. Although the likelihood of this, and other risks, would not diminish under a single organisation, their impact would be much greater, not least by affecting more students. Moreover, centralising the system, even if not nationalising it, would almost inevitably bring the delivery and responsibility of the service closer to the government's door. The experiences of Scotland (2000), New Zealand (2004) and the National Curriculum Tests crisis of 2008 in England all serve as a warning against centralisation, concentration of risk, and perceived political interference in assessment operations. For example, in writing about the three British examinations crises in 2000-2002, McCaig (2003) claimed that a common feature was government interference, particularly in too hastily imposing major reforms to the system.

5. RISKS AND BENEFITS OF OTHER APPROACHES

5.1. Alternative structural arrangements – other than a single, monolithic AB – are possible, but the main thread of this response is that not only do these not address the concerns expressed, they constitute different risks. Partitioning the industry by subject area or by operational function would, in the former instance, result in a reduction in choice and innovation whilst removing the incentive for better service, and in the latter case, risk a lack of operational coherence and continuity with attendant increased risks of catastrophic system failure.

5.2. Legitimate rises in grade outcomes will eventually reduce the ability of qualifications to discriminate. At GCE discrimination was improved with the introduction of the A* grade. Over time, more fundamental revision to the current grading model may be required to ensure that qualifications remain fit for purpose. However, examination grades are put to at least twenty different purposes (Newton, 2007), and McCaig (2003) showed that an issue underlying the 2002 GCE examination crises was successive governments' seeming unwillingness to define their expectations of the examination, a theme echoed in the Sykes Review (2010). Until this fundamental question is answered, fitness for purpose cannot be fully established and reorganising the structure of the ABs will simply prove to be a distraction, at best.

5.3. Some form of grade quota (norm-referencing) system could be introduced. This would prevent rising grade outcomes, but would fundamentally change the meaning of certification and impact on wider education policy such as school performance tables and targets. To ensure fairness (between years, specifications and awarding bodies), such a system ought to take into account the general ability of the cohorts of students entering qualifications (Stringer, in press). Detailed research into the implications of such a change would be required before widespread adoption.

5.4. Alternatively, a test-equating approach could be adopted (Wheadon, 2010), as per the KS2 National Curriculum tests or the Dutch national testing system. This is a very robust form of standards maintenance, empirically answering the question "What mark would a student who was awarded grade X on last year's examination need to have gained to be awarded grade X on this year's examination?" Whilst eliminating unjustified grade inflation, this approach does not, however, prevent rising outcomes, as the proportions attaining level 4 at KS2 demonstrate. In any case, the security risks associated with the necessary pre-testing for such high-stakes qualifications, allied to its cost and administrative logistics, currently render this option unfeasible.

6. CONCLUSION

6.1. In summary, this response argues the following.

6.1.1. There is no reason to believe that marking reliability would be improved by a change in AB structure. The marking reliability of GCSEs and GCEs is comparable with that of similar international assessments. Systematic improvements are likely to come from innovations such as on-line, question-level marking, and such innovation is more likely where competition between providers exists.

6.1.2. There is no evidence that rises in grade outcomes are due to either the existence of, or the competition between, several ABs. On the contrary, the awarding system is extremely tightly managed and many rigorous anticipatory/monitoring measures exist to ensure that inter-year and inter-AB standards are maintained.

6.1.3. It would be unwise to engineer a major organisational restructuring to create a solution to a problem that does not exist (inter-AB misalignment), whilst not solving a problem (unjustified rises in grade outcomes) that is perceived rather than real. The folly would be exacerbated because restructuring would not only introduce many new risks, but would jeopardise benefits in the current system, many of which are not obvious.

---

[i] "*The AB will appoint a single named person to be accountable directly to its governing body for ensuring the quality and standards of its qualifications."* Para 1.5. Statutory Code of Practice.

[ii] The following page of AQA's website provides a clear, yet comprehensive description of how grade awarding is undertaken: http://web.aqa.org.uk/over/stat_standards.php?id=03&prev=01

[iii] See: http://www.ofqual.gov.uk/downloads/category/163-enquiries-about-results

[iv] See Ahmed and Pollitt (2011) for a useful exploration of the relationship between assessment validity and mark scheme construction.

[v] See: http://www.ofqual.gov.uk/downloads/category/163-enquiries-about-results

[vi] The most strident such allegation was made by Mick Waters (former director of curriculum at the QCA), who claimed that *"the system is diseased, almost corrupt ... We've got a set of awarding bodies who are in a market place".* In his support, John Bangs (2010), former head of education at the NUT, opined that, *"I personally think there should just be a single examination board."* Bangs, J., MacBeath, J. and Galton, M. (2010) *Reinventing Schools, Reforming Teaching*. London: Routledge.

[vii] All the ABs, for example, have expert technical staff who are members of the Joint Council for Qualifications' (JCQ) Standards and Technical Advisory Group, and Ofqual's Standards and Technical Issues Group.

[viii] http://www.ofqual.gov.uk/news-and-announcements/130/745

[ix] Since 1973, CERP (and its predecessors) has undertaken 110 comparability studies of various types, the most recent one being a study of GCSE Law in 2011. All inter-AB comparability studies, invariably including a cross-moderation exercise of students' work, have been published.

[x] Joint Council for Qualifications protocol for the sharing of confidential data between awarding bodies.

[xi] How would the range of skills encompassed by GCSE English, for example, be validly assessed without extended writing?

[xii] In addition to the locally-developed CSE syllabuses, examples of such successful innovations range from the JMB's GCE General Studies syllabus to the Nuffield and Salters' science suites, the SMP mathematics developments, NEAB's 100% GCSE English coursework syllabus, and OCR's successful Computer Literacy and Information Technology (CLAIT) qualification. Doubtless there were failures too, but in the decades when these and other innovative qualifications were developed, a spirit of creativity prevailed.

[xiii] For example, the Nuffield Foundation 21st Century Science suite of GCSEs.

REFERENCES

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

Baird, J. (2007). Alternative conceptions of comparability. In P. E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.). Techniques for monitoring the comparability of examination standards. London: QCA.

Benton, T., & Lin, Y. (2011). Investigating the relationship between A level results and prior attainment at GCSE. NFER Report.

Boyle, B., & Charles, M. (2011). Re-defining assessment: the struggle to ensure a balance between accountability and comparability based on a 'testocracy' and the development of humanistic individuals through assessment. CADMO: An International Journal of Educational Research, 19(1), 55-65.

Bradshaw, J., Ager, R., Burge, B., & Wheater, R. (2010). Programme for International Student Assessment 2009: achievement of 15-year-olds in England. NFER Report.

Chamberlain, S., & Taylor, R. (2011). Online or face-to-face? An experimental study of examiner training. British Journal of Educational Technology, 42(4), 665-675.

Coe, R. (2007). Common Examinee Methods. In P. E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.). Techniques for monitoring the comparability of examination standards. London: QCA.

Cresswell, M. J. (1995). On Competition between Examining Boards. AEB Internal Report.

Daly, A., Baird, J., Chamberlain, S., & Meadows, M. (in press). Assessment reform: Students' and teachers' responses to the introduction of Stretch and Challenge at A-level. The Curriculum Journal, 23(1).

Fowles, D. (2005). Literature review on effects on assessment of e-marking. AQA Internal Report.

Fowles, D. (2009). How reliable is marking in GCSE English? English in Education, 43(1), 50-57.

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

Jones, B. E. (2009). Awarding GCSE and GCE - Time to Reform the Code of Practice? AQA Internal Report and presentation to CIEA conference, November 2010.

Malacova, E., & Bell, J. (2006). Changing Boards: investigating the effects of centres changing their specifications for English GCSE. The Curriculum Journal, 17(1), 27-35.

McCaig, C. (2003). School Exams: Leavers in Panic. Parliamentary Affairs, 56(3), 471-489.

Meadows, M., & Billington, L. (2005). A Review of Literature on Marking Reliability. Report produced for the National Assessment Agency.

Meadows, M., & Taylor, R. (2008). Enquiries about Results – an analysis of marking reviews. AQA Internal Report.

Newton, P. E. (2007). Evaluating assessment systems. London: QCA.

Newton, P. E., Baird, J., Goldstein, H., Patrick, H., & Tymms, P. (Eds.). (2007). Techniques for monitoring the comparability of examination standards. London: QCA.

Richmond, T., & Freeman, S. (2009). Rising marks, falling standards. Policy Exchange Report.

Pinot de Moira, A. (2009). Marking reliability & mark tolerances: Deriving business rules for the CMI+ marking of long answer questions. AQA Internal Report.

Stacey, G. (2011). Standard bearing: a new look at standards. Paper presented at the Ofqual A New Look At Standards event, London, UK.

Stringer, N. S. (In press). Setting and maintaining GCSE and GCE grading standards: the case for contextualised cohort-referencing. Research Papers in Education, 1-20.

Taylor, R. (2007). The impact of e-marking on enquiries after results. AQA Internal Report.

The Sir Richard Sykes Review. (2010). Retrieved from: http://www.conservatives.com/news/news_stories/2010/03/~/media/Files/Downloadable%20Files/Sir%20Richard%20Sykes_Review.ashx

Wheadon, C. B. (2011). An Item Response Theory Approach to the Maintenance of Standards in Public Examinations in England, (Doctoral Dissertation). Retrieved from http://etheses.dur.ac.uk/615/. University of Durham, Durham.