Examining assessment

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

AQA
Realising potential

Examining assessment

A compendium of abstracts taken from research conducted by AQA and predecessor bodies, published to mark the 40th anniversary of the AQA Research Committee

Edited by
Lena Gray, Claire Jackson and Lindsay Simmonds

Contributors
Ruth Johnson, Ben Jones, Faith Jones, Lesley Meyer, Debbie Miles, Hilary Nicholls, Anne Pinot de Moira and Martin Taylor

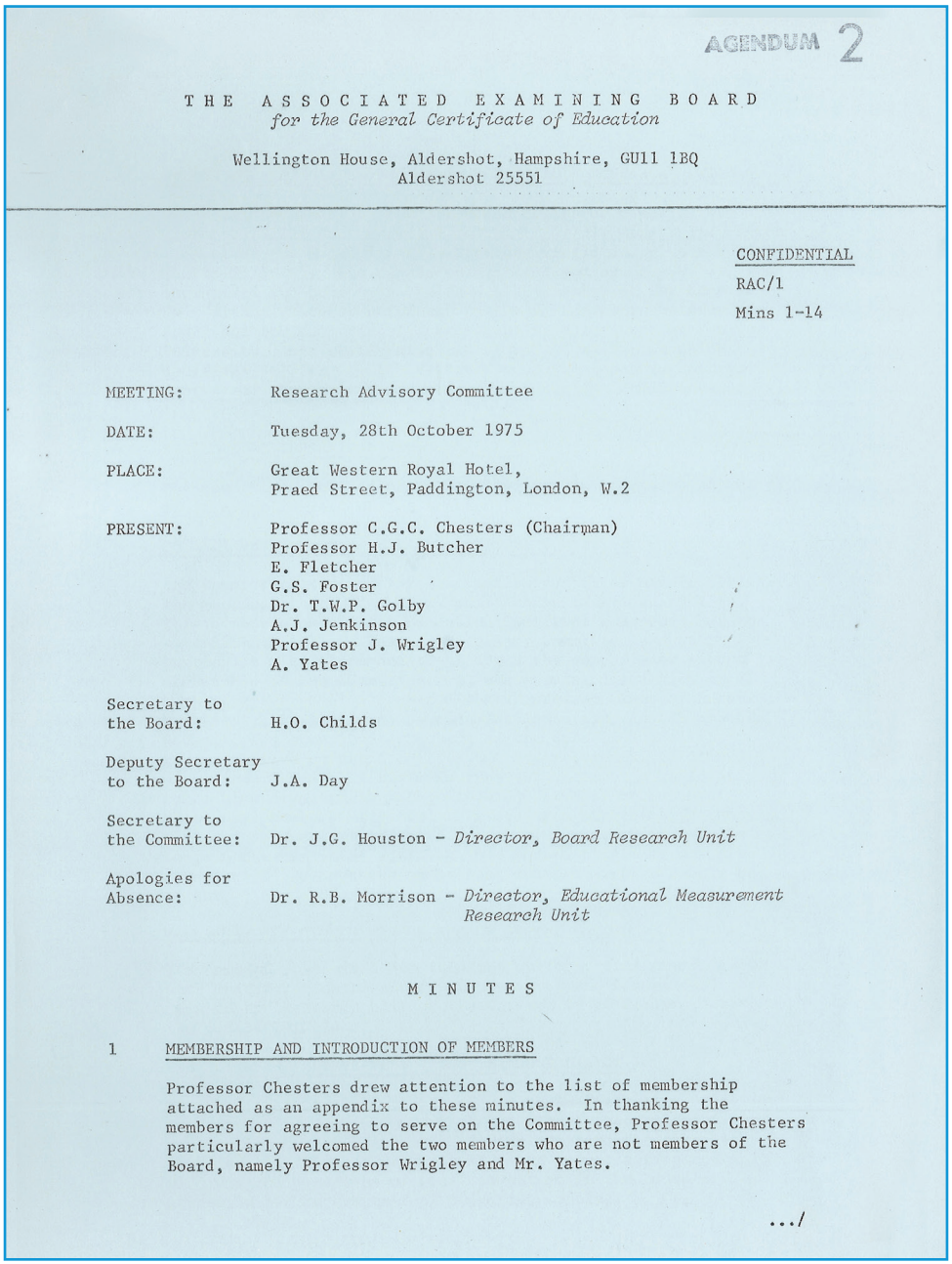
Director
Alex Scharaschkin

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

First published by the Centre for Education Research and Practice (CERP) in 2015.

Production by Rhinegold Publishing Ltd

Contents



AGENDUM 2

THE ASSOCIATED EXAMINING BOARD
for the General Certificate of Education

Wellington House, Aldershot, Hampshire, GU11 1BQ
Aldershot 25551

CONFIDENTIAL
RAC/1
Mins 1-14

MEETING: Research Advisory Committee

DATE: Tuesday, 28th October 1975

PLACE: Great Western Royal Hotel,
Praed Street, Paddington, London, W.2

PRESENT: Professor C.G.C. Chesters (Chairman)
Professor H.J. Butcher
E. Fletcher
G.S. Foster
Dr. T.W.P. Golby
A.J. Jenkinson
Professor J. Wrigley
A. Yates

Secretary to the Board: H.O. Childs

Deputy Secretary to the Board: J.A. Day

Secretary to the Committee: Dr. J.G. Houston - *Director, Board Research Unit*

Apologies for Absence: Dr. R.B. Morrison - *Director, Educational Measurement Research Unit*

MINUTES

1 MEMBERSHIP AND INTRODUCTION OF MEMBERS

Professor Chesters drew attention to the list of membership attached as an appendix to these minutes. In thanking the members for agreeing to serve on the Committee, Professor Chesters particularly welcomed the two members who are not members of the Board, namely Professor Wrigley and Mr. Yates.

.../

Minutes from the AEB's first ever research committee meeting that took place on 28 October 1975 at the Great Western Royal Hotel in London

Introduction

The education system is continuously evolving: in fact, the only constant feature is change. High-stakes examinations loom large in this paradigm – but these must flex and adapt.

This book, created to mark the 40th anniversary of the AQA Research Committee, follows on from Mike Cresswell's 1999 publication *Research Studies in Public Examining*, which was produced just before the Associated Examining Board (AEB) and Northern Examinations and Assessment Board (NEAB) merged to form a single awarding body – the Assessment and Qualifications Alliance (AQA). I am indebted to Mike for this valuable title; many of the abstracts from the work completed by the AEB during the 80s and 90s are reproduced here.

This volume also features examples of research that AQA has carried out more recently, within its Centre for Education Research and Practice (CERP). Standards and awarding continue to be integral research topics, however, during the last decade, our attention has turned to marking, validity and assessment design. We are also considering the impact that assessment has on students and stakeholders, and how to ensure that non-examination assessments are fair.

We have curated this collection along these broad themes. This is a snapshot in time of our research; as we move to adopt new techniques, so our studies will diversify. Most of the papers cited here can be read in full on our website at cerp.org.uk.

This is an exciting and challenging time to work in assessment, and research is undertaken against a backdrop of lively discussion. AQA's research is influenced by changes in policy, although we use our research to inform and advise, too. There is much work to be done, but as this compendium illustrates, we have come a long way. May I take this opportunity to thank our researchers – past and present – for their significant contributions to both the AQA Research Committee and CERP.

Alex Scharaschkin
Director, Centre for Education Research and Practice (CERP)

A brief history of research conducted by AQA and predecessor bodies

Notes from the North

By the middle of the 20th century, assessment research within disparate northern awarding organisations had become systematic, and in 1992 the Northern Examinations and Assessment Board (NEAB) was born. *Ben Jones*, CERP's Head of Standards, charts the group's evolution from fledgling research units to a centre of excellence, and highlights key moments in its output

Early years

It is difficult to identify when the northern examination boards officially started research work, largely because the definition of that activity is fluid. However, research within the Joint Matriculation Board (JMB) – the largest of the northern awarding organisations – can be traced back to the mid 1950s.

Dr J. A. Petch, secretary to the JMB during 1948-65, had a lively interest in research and encouraged its development. Projects were led by Professor R. A. C. Oliver, a JMB member who represented the Victoria University of Manchester. Oliver launched the aptly titled Occasional Publications ('OPs'). The first of these, OP1, was entitled *A General Paper in the General Certificate of Education Examination* and was published in July 1954.

By mid 1960s, JMB research had become more strategic. The 1964 JMB annual report (p. 7) states that:

'The series [of OPs] represents some of the fruits of investigations and researches which have been carried out for a number of years. The Board has now decided that the time is opportune to make more formal provision for this kind of work. At its meeting in August it approved a proposal, foreshadowed in the previous year's report, that a Research Unit be established with its own staff and necessary working facilities. The present Secretary to the Board was appointed the first Director of the Unit.'

Petch's appointment to the newly created post of director of research, to manage the work of the Research Unit (RU), would greatly enhance the quality of the board's activities. Gerry Forrest was appointed as Petch's replacement when the latter retired in 1967, a post Forrest was to hold until December 1989.

In 1965, the board established the first committee that would oversee its research. This was succeeded by the Research Advisory Committee (RAC), which operated from 1973 to 1992. Professors Jack Allanson and Tom Christie were long-standing and active members of the RAC, and, respectively, its chairs. Besides being eminent professors from two of the JMB's constituent universities, both were acknowledged national leaders in educational assessment. They were members of the Department of Education and Science's (DES) Task Group on Assessment and Testing (TGAT), and co-authors of the influential TGAT Reports (1987; 1988).

One of Christie and Forrest's collaborative exercises culminated in the seminal *Defining Public Examination Standards* (1981); it was rumoured that Sir Keith Joseph carried this publication with him when he was Secretary of State for Education and Science (1981-86).

The two longest-serving members of the RU staff over this period were Austin Fearnley (who worked for the unit from 1971 until 2006) and Dee Fowles (who contributed during the period 1979-2009). Both produced many research reports and papers, some of which are referred to below. (Pre-AQA, individual staff members were not identified as authors of internal papers.)

Review of work undertaken 1971–2000

Projects undertaken by the JMB over this period bear a likeness to contemporary assessment research work.

As early as 1953, for example, concerns were expressed over the standard of candidates' spoken English. The controversy continues – in England at least – and the speaking and listening component has recently been decoupled from the GCSE qualification grade, so that it now exists as a three-level, internally assessed endorsement. The following extract from the JMB's annual report for 1954 evokes recent discussions:

'There is at present much criticism current of the inability of some school pupils to use their mother tongue correctly, whether in writing or orally. It is not a new source of complaint but possibly the general standard of spoken English is at all events not rising. In 1952 the Board was requested by one school to conduct an experiment in testing some of its pupils in spoken English. In 1953 pupils from 5 schools were tested; in 1954 … 59 were selected to give as wide a spread as possible of type and region and 1,775 candidates were examined. The experiment is to be continued in 1955. *The oral test in English is completely dissociated from the Examination for the General Certificate.*' [emphasis added] (p. 9)

Plus ça change, plus c'est la même chose!

Coursework and moderation

Internally assessed components often differ in content, structure and regulation from the coursework components of last century. Many comprise controlled assessments, which are governed by subject-specific requirements. In future, the standard title of non-examined assessment will be used as a self-explanatory umbrella term (see pp. 42–45). Nevertheless, today's issues – primarily regarding manageable, effective and fair moderation procedures – are very similar to those faced by the JMB during the 1970s, when coursework was becoming increasingly popular. This was evident in the JMB's GCE English O-level Syllabus D, which would eventually transmute into the NEAB's 100 per cent coursework GCSE English specification. Much research into methods of moderation was undertaken as a consequence of these developments.

OP38: *JMB experience of the moderation of internal assessments* (1978) reviewed different inspection and statistical approaches to moderation taken from JMB experience in GCE and trial O-level and CSE examinations. The mean marks of candidates in each centre were calculated separately for the moderating instrument and for the internally assessed component. The mean marks were scaled to a common maximum mark and the difference between them compared. If this difference was outside pre-determined tolerance limits, a flat-rate adjustment was applied to the internally assessed marks.

However, statistical moderation was not without its critics. Various refinements to the JMB's standard procedure were introduced over the years. For example, teaching sets within a large-entry centre could be moderated separately, although centres were always encouraged to standardise their own assessments. The variations culminated in a rather elaborate procedure for the moderation of project work assessments in A-level Geography, which operated for the first time in 1987. It combined inspection of samples of work with the standard statistical method, and took account of the correlation between a centre's moderating instrument and project marks. When this fell below an acceptable level, a team of moderators could override the statistical outcome on the basis of the sample of coursework that all centres were required to submit with their marks. The moderators were not confined to flat-rate adjustments when they reviewed any of the statistically derived adjustments, but they were required to retain the candidates' rank order as established by the teachers' marks.

(By an extraordinary coincidence, the Joint Council for Qualifications (JCQ) is currently making routine a common analysis – very similar to that described above – to identify centres in which the differences between their internally and externally assessed marks appear anomalous.)

Objective test questions (OTQ)

The use of Objective Test Questions (OTQs) gathered pace during the 1960s. The GCE General Studies specification – which in its heyday attracted over 40,000 entries – comprised 60 per cent OTQs. Other specifications, notably GCE Economics, Geology, Physics and Chemistry, also had substantial OTQ components. The RU undertook various research projects both to inform the design of OTQ tests, and to ensure that standards were maintained.

By the 1970s, pre-testing of OTQs – by means of the large-scale recruitment of centres to deliver a balanced pre-test population and valid item statistics – was acknowledged as being very demanding on centres and exam board staff. An alternative method, which was investigated by the RU in 1974, involved asking a group of item writers to predict the suitability of items and the level of performance on each that would be expected from the candidature. AQA revived this method (known as the Angoff procedure) years later, with facility predictions for candidates at the key grade boundaries averaged and summed to give suggested grade boundaries for the OTQ component.

In 1974, Dee Fowles and Alan Willmott published a useful introductory guide to Rasch modelling for objective test items entitled *The Objective Interpretation of Test Performance; the Rasch Model applied* (NFER, 1974) However, nothing came to fruition with the Rasch approach in that decade – perhaps the model and the largely opaque data processing did not breed confidence – but AQA is currently applying it in a few contexts, for example equating inter-tier standards at GCSE Grade C.

Curriculum reforms

At the time of writing, awarding organisations and Ofqual are preparing for the biggest reform of general qualifications in a generation. GCSEs are returning to a linear structure and will adopt a numerical nine-point grade scale. GCEs are also becoming linear, with the AS qualification being decoupled from A-level. Such changes are not unprecedented.

The JMB began investigating a unified examination system – to replace GCE O-level and the CSE – as early as 1973 (15 years before the advent of the GCSE). The RU was involved in preparatory work with four CSE boards, and the 1976 annual report noted that:

'The Research Unit was responsible for the preparation and detailed analyses of the data for all the 15 studies in which the JMB is involved and, in addition, prepared the statistical sections of the reports submitted to the Schools Council for 10 of the 15 subjects. The staff of the Unit are also consulted by the 16+ Working Parties on matters of assessment and provide reports and undertake investigations when required.'

The pilot joint examinations of the 1970s brought together CSE and O-level standards relatively painlessly, with the two groups of examiners able to negotiate the grade boundaries for the overlapping grades against exemplars from the parallel CSE and GCE examinations.

The RU became involved, in collaboration with the Schools Council, in two important projects. The first used the experience of graded objective schemes in the linear subjects French and Mathematics; it asked examiners to build up grade descriptions in these subjects that could inform awarding (Bardell, Fearnley and Fowles, *The Contribution of Graded Objective Schemes in Mathematics and French*, JMB 1984). In the second, examiners explored the relationship between the grades and the assessment objectives of the joint examinations in History and English (Orr and Forrest, 1984, see p. 60), while Physics and English examiners scrutinised scripts and attempted to describe the performance of candidates at the key grades (*Grade characteristics in English and Physics*, Forrest and Orr, JMB 1984). Subsequently, GCSE grade criteria for the nine subjects were developed and passed on in 1986 for use by the examining bodies.

However, by the time the GCSE was introduced (first examination in 1988), any thoughts of strict criterion referencing had been abandoned. It was recognised that a compensatory, judgemental grading approach, supported with – what now seems to be rudimentary – statistical support evidence would be needed. Thus, the GCSE enjoyed a fairly uneventful launch.

In the 1990s, awarding organisations grappled with the introduction of a secondary GCE award, known as the Advanced Supplementary award (these were designed to be of the same standard as Advanced Level, but covering approximately half the subject content). The RAC had begun recommending, and conducting investigations into, an intermediate GCE qualification, more akin to the Advanced Subsidiary award (intended to be half the content of a full A-level award but at a lower standard, i.e. what an A-level student might be expected to achieve after one year's study), which was eventually adopted in 2001. Therefore, it could be said that the RAC played something of a prophetic role in arguing for the efficacy of the latter design.

Comparability of standards

One of the main features of the 1970-2000 period was the development of standard setting and maintenance, particularly the various aspects of comparability. Researchers examined the traditional dimensions of comparability – inter-board, inter-year, inter-subject – as well as topics that have contemporary significance, e.g. establishing and describing the standards of a new grade scale and ensuring comparability of optional routes within the same qualification.

The introduction of the National Curriculum and its concomitant Key Stage test scores, together with the creation of longitudinal national matched datasets for individual students, allowed more sophisticated, valid and reliable statistical modelling of subject outcomes to be made. Additionally, concurrent research, particularly in the Associated Examining Board (AEB) in the 1990s, indicated that even experienced awarders' judgements were subject to unreliability and bias. The most notable example was the effect on examiners' judgement of student performance by differences in question paper/mark scheme demand. This was identified by Mike Cresswell and Frances Good, and gave rise to the 'Good & Cresswell effect' (summarised as the tendency of examiners to compensate insufficiently for variation in question paper/mark scheme demand when deciding on grade boundaries). Cresswell was Head of Research at the AEB – and subsequently AQA – from 1991 to 2004, after which he was appointed CEO of AQA until his retirement in 2010.

Grade awarding has been increasingly guided by statistical predictions. The application of this approach, via national subject prediction matrices derived from reference years at the specification's inception, means that variation in both inter-board and inter-year standards are – by this definition – now discounted by the awarding method itself.

In recent years, and certainly in the era of comparable outcomes, inter-subject standards has generally been considered too complicated an issue, both philosophically and methodologically, to devote substantial research resources to. There have been exceptions to this rule. For example, as a result of a judgemental and statistical research exercise by AQA, Ofqual recently endorsed a gradual readjustment of standards in the former's GCSE Dance award to align it more closely with GCSE Drama. However, in the 1970s JMB pioneered the method of subject pairs, which was designed to identify syllabuses that appeared to be relatively leniently or severely awarded. The routine analyses were undertaken annually and comprised one of several statistical inputs – albeit a secondary one – to awarding meetings. ■

Strength in numbers

The NEAB eventually joined forces with the Associated Examining Board (AEB) to form the Assessment and Qualifications Alliance (AQA). The two organisations officially merged in 2000, having amassed an impressive body of assessment research. This provided a sound basis for the work subsequently carried out within the Centre for Education Research and Practice (CERP). Senior researcher *Martin Taylor* reflects on the development of research activities throughout the period 1998-2015, with reference to the AEB's achievements in the main

Like the NEAB, the AEB had a long tradition of high-quality assessment research. Dr Jim Houston led the AEB Research and Statistics Group from its inception in 1975 until 1991, when he was succeeded by Mike Cresswell (see above). Research was completed under the guidance of the AEB Research Advisory Committee, and research topics were developed in line with education policy.

Cresswell's 1999 publication *Research Studies in Public Examining* highlights the varied research achievements of the AEB during 1975-1999, and for that reason, discussion of the AEB's illustrious history will remain brief here.

The introduction of school performance tables fundamentally altered the way in which results were interpreted and led to ever-greater public scrutiny. Throughout the late 90s, the AEB formed an alliance (AQA) with the NEAB and City and Guilds (CGLI). By the turn of the century, the alliance evolved into a merger (this excluded CGLI), and a single awarding organisation was formed. The AQA Research Committee replaced the separate AEB and NEAB advisory committees. In 2011, the research department was rebranded as the Centre for Education Research and Policy (CERP), later renamed the Centre for Education Research and Practice (2014). For simplicity, the abbreviation 'CERP' will be used below to describe all research work since 1998.

Throughout this period, CERP's output can be broadly divided into two areas: statistical and support work, and general research. Examples of work in the first area include generating results statistics for AQA and the JCQ; advising on moderation of internal assessment; awarding; and supporting specification development by ensuring that new specifications and assessments are technically sound. Most of the work described here falls into the second area.

General research is not necessarily connected to immediate operational issues or specific examinations, but provides important background knowledge for the general improvement of assessments and procedures. In recent years, AQA has been keen to enhance its reputation for expertise in assessment, and to develop its credentials for speaking authoritatively to the regulators, government and the wider public about the current examination system, and about assessment more generally.

Maintaining standards

Improving techniques for the establishment and maintenance of standards has been a constant theme since 1998.

Initially, these techniques included delta analysis (whereby comparability across awarding bodies and between years was monitored on the

assumption that results within each centre type should be similar); common centres analysis (whereby results for centres entering candidates for a subject in successive years were expected to be similar); subject pairs (whereby candidates entering two subjects were expected to obtain similar results in those subjects); and judgemental methods.

When the Curriculum 2000 modular AS and A-level qualifications were first certificated in 2001 and 2002, an approach to standard setting that relied mainly on judgement would have been untenable, as the structure of the qualifications was very different from that in the previous linear syllabuses. Therefore, predicted outcomes were used for the first time, alongside expert judgement. The predictions sought to carry forward standards from the previous syllabuses at a national level within each subject, taking account of candidates' prior attainment as measured by their average GCSE scores from one or two years earlier. The philosophy underpinning this approach (which is now seen in Ofqual's comparable outcomes policy) is that, in general, candidates with a particular prior attainment should gain the same A-level grade as their counterparts in previous years.

The approach for generating predicted outcomes was also used for inter-awarding body statistical screening: a process instigated in 2004 by the JCQ Standards and Technical Advisory Group (STAG). It consisted of a post-hoc statistical review of the previous summer's GCSE, AS and A-level results. Actual results in each specification were compared with the predicted results, which were calculated from the national results in the subject in question, taking account of the entry profile for the individual specification. 'Entry profile' means prior attainment (in the case of AS and A-level) or current attainment (in the case of GCSE), measured by candidates' average GCSE scores. At the time, predicted outcomes for GCSE awards were not generally used, and statistical screening was an important way of checking whether standards were comparable across all specifications in a subject. If any deviation was found, an appropriate adjustment to the following year's award was normally applied (unless further investigation revealed a justifiable reason for the deviation).

In the past, comparability studies played a significant role in all examination boards' research departments (as outlined above). By 1998, statistical techniques were gaining importance, and the use of regular, large-scale judgemental exercises soon ceased.

Collaborative projects

Until the early 2000s, promotion of individual exam-board syllabuses was carried out in a fairly discreet manner. AQA did not have a marketing department; when new syllabuses were being devised, CERP often carried

out surveys of centres to investigate teachers' preferences in relation to aspects that were not specified by the regulators. These surveys were generally conducted by post, but telephone studies and focus groups became increasingly common.

A significant part of CERP's work in the early 2000s was associated with the World Class Arena: an initiative led by the Department for Education and Skills and the Qualifications and Curriculum Authority (QCA) to improve education for gifted and talented students, especially in disadvantaged areas of the country. AQA had a contract with QCA to administer, market and evaluate World Class Tests for pupils aged nine and thirteen, in maths and problem solving. The research required by the project included analyses of the technical adequacy of the tests, provision of data to underpin the standard setting processes, and review of the results data.

In summer 2001, a report was published detailing a study that AQA had undertaken on behalf of the JCQ. The report included: a review of past policy and practice on differentiation; an investigation of the incidence of 'falling off' the higher tier or being 'capped' on the foundation tier; a summary of the views of teachers, examiners and students; and an analysis of the advantages and disadvantages of various forms of differentiation.

E-marking

In the early 2000s, CERP carried out extensive e-marking research, which included: trialling, investigating reliability, evaluating the impact on enquiries about results, and consideration of the extension to long-form answers. Gains in reliability from using item-level marking (compared to the traditional method of sending a whole script to a single examiner) were investigated. Work was also carried out to compare the reliability of online and face-to-face training. More generally, reliability of marking has been a constant theme for CERP, and recent research has focused on levels-based mark schemes, which are commonly used for extended-response questions.

Sharing our expertise

Soon after the introduction of Curriculum 2000, QCA instigated a series of annual technical seminars, which were intended to address the numerous issues arising from modular examinations and from the greater emphasis on the use of statistics in awarding. These seminars have continued under the auspices of Ofqual, although the title and focus have recently changed. From the outset, members of CERP have played a major role in presenting items at these seminars.

From December 2003, CERP was involved in the work of the Assessment Technical Advisory Group, which had been set up to support the Working

Group on 14-19 Reform, chaired by Mike Tomlinson. The purpose was to develop and advise on models of assessment to support the design features of the working group's proposed diploma model. The working group's proposals were published in October 2004 but were rejected by the government; instead, the 2005 Education and Skills White Paper announced a set of Diploma qualifications, covering each occupational sector of the economy, to run alongside GCEs and GCSEs. CERP convened a project group that produced recommendations (presented to QCA in early 2007) on how these new Diplomas should be graded.

Expanding themes

CERP's general research has understandably tended to focus on assessment issues, but broader educational themes have also been considered from time to time. Recent research has included: validity theory; university entrance worldwide; and analysis of educational reforms as they relate to a 'choice and competition model' of public provision. CERP's current aim is to continue to carry out and disseminate high-quality assessment research; the findings of which will help AQA to produce assessments that fairly test students, are trusted by teachers and users of qualifications, and are of the highest technical quality. CERP defines its work in four major areas:

Awarding, standards and comparability emphasises CERP's central role in ensuring that grading standards are maintained.

Assessment quality refers to the need to design assessments and mark schemes that are valid, fair and reliable.

Exam statistics, delivery and process management is about providing and maintaining examination statistics and supporting materials, and giving technical support to the development of procedures such as standardisation and moderation.

Innovation in assessment design and delivery involves improving current processes through the use of evidence-based design, and boosting validity and reliability through alternative forms of assessment and marking models. ■

The following collection of abstracts offers a summary of the work undertaken by AQA and predecessor bodies during 1975-2015. Many of these papers are available in full at cerp.org.uk.

Standards and comparability

Defining the term 'standard' in the educational context is fraught with difficulties. Interpreting what is written on an examination script introduces subjectivity. Further, attainment in education is an intricate blend of knowledge, skills and understanding, not all of which are assessed on any one occasion, nor exemplified in any one single script. From year to year, the question difficulty and the demand of question papers will be different. Therefore, when two scripts are compared, the comparison cannot be direct. The standard of each script has to be inferred by the reader, and each inference is dependent on interpretation. Different individuals will place different values on the various aspects of the assessment, and so conclude different things from the same student performance.

Alongside the difficulty in defining standards in relation to education, there is also confusion about the way the term is interpreted and used. Public examination results are used in a variety of different ways, many of which exceed the remit of the current examining system. Fundamentally, the problem stems from the need to distinguish between the standards of the assessment (i.e. the demand of the examination) and the standards of student attainment (i.e. how well candidates perform in the examination).

Defining the standard for an examination in a particular subject involves two things: firstly, we have to establish precisely what the examination is supposed to assess; secondly, since standards represented by the same grade from examinations of the same type (GCSE, for example) should be comparable, we have to establish (at each grade) what level of attainment in this subject is comparable to that in other examinations of the same type. The need for fairness means that comparability of standards set by different awarding organisations, in different subjects, and across years, is a key focus. Over the years, work has focused on the methodological aspects of comparability studies, both from a statistical and judgemental perspective. New statistical methods of investigating comparability of standards have been increasingly advocated and developed, as indicated by the selection of reports that follow.

Comparability in GCE: A review of the boards' studies 1964-1977

Bardell, G. S., Forrest, G. M. and Shoesmith, D. J. (1978)

This booklet is concerned with the inter-board studies, undertaken since 1964, to compare grading standards in the ordinary A-level examinations

of two or more GCE boards. The booklet is divided into five sections. The first provides background to the studies and describes the differences that exist between the nine GCE boards in the UK, such as clientele, syllabuses and examinations. Each of the sections 2, 3 and 4 are based on one of the three major approaches to monitoring inter-board grading standards that have been used in recent years: analysis of comparability using examination results alone, monitor tests and cross-moderation.

The first section draws attention to reported differences in examination results, leaving tacit the numerous similarities that are reported. The second explores the limitations and caveats that regularly accompany comparability using reference tests. The third explores the difficulty of determining which is the correct standard when studies indicate that two or more boards differ in standard.

The conclusion summarises the lessons to be learnt from the GCE experience in monitoring grading standards over the decade. It is concluded that a degree of error in public examinations is currently unavoidable. Differences between the boards could be resolved through the introduction of a national curriculum. However, this is unlikely to receive much support – particularly from teachers, who value the flexibility of the British system.

Defining public examination standards

Christie, T. and Forrest, G. M. (1981)

This study seeks to explore the nature of the judgement that is required when examination boards are charged with the responsibility of maintaining standards. The argument is generalisable to any public examination structure designed to measure educational achievement, although the current focus is on the A-level procedures of the Joint Matriculation Board (JMB). Historical definitions of standards stress the importance of maintaining a state of equilibrium in examination practice, between attainment by reference to a syllabus and attainment by reference to the performance of other candidates. Present practice in the JMB is reviewed to see how this required equilibrium is maintained in the examiners' final meetings and, on the basis of an analysis of JMB statistics, it is concluded that the demands of comparability of standards between subjects and within a subject have diverged over time. A contest model of grading of the implementation of standards is adduced.

Two theoretical models of grading are then considered from the point of view of how well they fit to models of the nature of education achievement. A third model – limen-reference assessment – is derived, which is thought to represent current practice in public examining boards; its properties and

potential development are discussed. There appears to be no compelling theoretical reason for adopting any one of these models. Finally, the differing benefits of the approaches – emphasising either parity between subjects or parity between years – are briefly reviewed in the context of the responsibility of a public examination system; namely, the provision of feedback to selectors, pupils, subject teachers and the wider society. In view of the imminent changes in certification at 16+, and the continuing problems of sixth-form examinations, it is hoped that this study will outline the priorities that should guide public examination boards in maintaining standards.

Norm and criterion referencing in public examinations

Cresswell, M. J. (1983)

Neither traditional norm-referencing nor traditional criterion-referencing techniques can be applied to public examining. However, elaborations of these techniques can be seen to offer potential solutions to the problem of imposing comparable standards through the grading schemes of different examinations. The choice between an empirical or judgemental definition of equivalence of performance standards, and hence a normative or criterion-related grading scheme, is primarily a value judgement.

Most examination boards currently attempt to use both approaches, and where they produce similar results, this can be reassuring. However, since the two approaches are based upon quite different conceptions of what constitutes equivalence of performance, when they produce different results no accommodation between them is possible. In these circumstances, the emphasis given to one, rather than the other, is again a value judgement.

A comparability study in A-level Physics: A study based on the summer 1994 and 1990 examinations

Northern Examinations and Assessment Board on behalf of the Standing Research Advisory Committee of the GCE Boards

Fowles, D. E. (1995)

The report describes the conduct and main findings of the 1994 inter-board comparability study of A-level Physics examinations. The design of the study required each board to provide complete sets of candidates' work for its major syllabus at each of the A/B, B/C and E/N grade boundaries. In addition, four boards were selected for comparison of the 1990 and 1994 syllabuses and scripts. Each board nominated two senior examiners to act as scrutineers in the study. The study comprised three strands: a statistical analysis of the examination results, a syllabus review and a cross-

moderation exercise. The examination statistics suggest relative leniency in grading on the part of the WJEC at the grade A/B and B/C boundaries and of OCSEB (Nuffield) at the grade E/N boundary. The syllabus review required the scrutineers to rate the relative demands made by each syllabus (using syllabus booklets, question papers, mark schemes and other support materials) against an agreed set of factors. Four factors were identified for Physics: content, skills and processes, structures and manageability of the question papers; and practical skills.

The results of the cross-moderation exercise suggested that, at the grade A/B and B/C boundaries, three of the 1994 syllabuses – those of the AEB, UCLES and the WJEC – were relatively leniently graded. Scrutineers were generally satisfied with the methodology, and found the study a useful means of evaluating their own work in relation to that of the other boards. However, many noted that the exercise involved making holistic judgements, whereas current awarding practice involves making separate judgements on each component. They also pointed out that the products they were asked to compare were rather different in nature, despite sharing the title 'physics'.

On competition between examining boards
Cresswell, M. J. (1995)

This paper uses game theory to analyse the consequences of competition in terms of standards between examining boards. The competitive relationship between examining boards is shown to have elements of a well-known paradox: the prisoners' dilemma. It is also demonstrated in the paper that, even if only reasons of narrow self-interest are considered, examining boards should not compete by reducing the standards represented by the grades that they issue. It is also shown that a rational, but purely self-interested, examining board would not compete in this way even if it felt that the chances of its actions being detected by the regulators were small. Finally, it is argued that a rational self-interested examining board would not compete on standards even if another board chose to do so. Furthermore, it is claimed that the board would correct its own standards if, through error, they were lenient on a particular occasion.

Defining, setting and maintaining standards in curriculum-embedded examinations: judgemental and statistical approaches

(pp. 57–84 in *Assessment: Problems, Developments and Statistical Issues*, edited by H. Goldstein and T. Lewis; [Chichester, Wiley])
Cresswell, M. J. (1996)

This paper analyses the problems of defining, setting and maintaining standards in curriculum-embedded public examinations. It argues that the setting of standards is a process of value judgement, and shows how this perspective explains why successive recent attempts to set examination standards solely on the basis of explicit written criteria have failed, and, indeed, were doomed to failure. The analysis provides, for the first time, a coherent theoretical perspective that can be used to define comparable standards in quite different subjects or assessment domains. The paper also reviews standard-setting methods in general, and statistical approaches to establishing comparable examination standards, in particular. It explores in detail the various assumptions that these approaches make. The general principles underlying the analysis in the paper apply equally well to other means and purposes of assessment, from competence-based performance assessments to multiple-choice standardised tests.

The comparability of different subjects in public examinations: A theoretical and practical critique

(*Oxford Review of Education*, Vol. 22, No. 4, 1996, pp. 435–442)
Goldstein, H. and Cresswell, M. J. (1996)

Comparability between different public examinations in the same subject – and also different subjects – has been a continuing requirement in the UK. There is a current renewed interest in between-subject comparability, especially at A-level. This paper examines the assumptions behind attempts to achieve comparability by statistical means, and explores the educational implications of some of the procedures that have been advocated. Some implications for examination policy are also briefly discussed.

Examining standards over time

(*Research Papers in Education* Vol. 12, No. 3, 1997, pp. 227–247)
Newton, P. E. (1997)

Public examination results are used in a variety of ways, and the ways in which they are used dictate the demands that society makes of them. Unfortunately, some of the uses to which British examination results are

currently being put make unrealistic demands. The government deems it necessary to measure the progress of 'educational standards' across decades, and assumes that this can be achieved to some extent with reference to pass rates from public examinations; hence, it demands that precisely the same examining standards must be applied from one year to the next. Recently, it has been suggested that this demand is not being met and, as a consequence, changes in pass rates may give us a misleading picture of changing 'educational standards'. Unfortunately, this criticism is ill-founded and misrepresents the nature of examining standards, which, if they are to be of any use at all, must be dynamic and relative to specific moments in time. Thus, the notion of 'applying the same standard' becomes more and more meaningless the further apart the comparison years. While, to some, this may seem shocking, the triviality of the conclusion is apparent when the following are borne in mind: (a) the attempt to measure 'educational standards' over time is not feasible anyway; (b) the primary selective function of examination results is not affected by the application of dynamic examining standards.

Statistical analyses of inter-board examination standards: better measures of the unquantifiable?

Baird, J. and Jones, B. (1998)

Statistical analyses of inter-board examination standards were carried out using three methods: ordinary least squares regression, linear multilevel modelling, and ordered logistic multilevel modelling. Substantively different results were found in the candidate-level regression compared with the multilevel analyses. It is argued that ordered logistic multilevel modelling is the most appropriate of the three forms of statistical analysis for comparability studies that use the examination grade as the dependent variable. Although ordered logistic multilevel modelling is considered an important methodological advance on previous statistical comparability methods, it will not overcome fundamental problems in any statistical analysis of examination standards. It is argued that, ultimately, examination standards cannot be measured statistically because they are inextricably bound up with the characteristics of the examinations themselves, and the characteristics of the students who sit the examinations.

Would the real gold standard please step forward?

(*Research Papers in Education*, Vol. 15, No. 2, 2000, pp. 213–229)
Baird, J., Cresswell, M. J. and Newton, P. E. (2000)

Debate about public examination standards has been a consistent feature of educational assessment in Britain over the past few decades. The most frequently voiced concern has been that public examination standards have fallen over the years; for example, the so-called A-level 'gold standard' may be slipping. In this paper, we consider some of the claims that have been made about falling standards, and argue that they reveal a variety of underlying assumptions about the nature of examination standards and what it means to maintain them. We argue that, because people disagree about these fundamental matters, examination standards can never be maintained to everyone's satisfaction. We consider the practical implications of the various coexisting definitions of examination standards and their implications for the perceived fairness of the examinations. We raise the question of whether the adoption of a single definition of examination standards would be desirable in practice, but conclude that it would not. It follows that examining boards can legitimately be required to defend their maintenance of standards against challenges from a range of possibly conflicting perspectives. This makes it essential for the boards to be open about the problematic nature of examination standards and the processes by which they are determined.

A review of models for maintaining and monitoring GCSE and GCE standards over time

Cresswell, M. J. and Baird, J. (2000)

Maintaining and monitoring GCSE/GCE examination standards involves comparing the attainment of students taking examinations on different occasions. When the standards of a particular grade are maintained, these comparisons are made with a view to identifying the level of performance on the new examination that represents attainment of the same quality as work that received that grade in the previous examination on the same syllabus.

Monitoring involves comparing work that has already been awarded the same grade to see if the performances of the candidates for both examinations represent attainment of equal quality and, if not, to estimate the direction and size of any difference. The procedures used to maintain and monitor GCSE/GCE standards involve both professional judgement and statistical data.

Are examination standards all in the head? Experiments with examiners' judgements of standards in A-level examinations

(*Research in Education*, Vol. 64, 2000, pp. 91–100)
Baird, J. (2000)

Examination grading decisions are commonplace in our education system, and many of them have a substantial impact upon candidates' lives – yet little is known about the decision-making processes involved in judging standards. In A-level examinations, judgements of standards are detached from the marking process. Candidates' work is marked according to a marking scheme and then grade boundary marks are judged on each examination paper, to set the standard for that examination. Thus, the marking process is fairly well specified, since the marking scheme makes explicit most of the features of candidates' work that are creditworthy. Judging standards is more difficult than marking because standards are intended to be independent of the difficulty of the particular examination paper. That is, candidates who sit the examination in one year should have the same standard applied to their work as those who sat the examinations in previous years (even though the marks may differ, the grade boundaries should compensate for any changes in the difficulty of the examination). Note that if the marking and standards-judgement tasks are not detached, and grading is done directly, the problems inherent in standards judgements are still present – although they may not be as obvious to the decision maker.

Subject pairs over time: A review of the evidence and the issues

Jones, B. (2003)

It is incumbent on the awarding bodies in England, Wales and Northern Ireland to aim to ensure that their standards are equivalent between different specifications and subjects, as well as over time; although the regulatory authorities do not stipulate exactly what is meant by this requirement, nor how it should be determined. Until relatively recently, subject pairs data comprised one of several indicators that informed awarders' judgemental boundary decisions. The last decade has seen a demise in the use of this method due to the assumptions associated with it being seriously

undermined. This paper summarises the main literature from this period that argued against the validity of the method. It then presents and discusses GCE subject pairs results data for the last 28 years of the JMB/NEAB – one of the GCE boards that used the subject pairs method most extensively. Finally, it is noted that many of the issues associated with the subject pairs method have their roots in whether grade awarding, and grading standards, are intended to reflect candidate ability or attainment. Although the emphasis is currently on the latter, it is noted that this is largely a phenomenon of the last 30 years or so. Were the balance to move back towards the equating of standards with ability, then the subject pairs method, or something similar, might – in certain situations (e.g. equating cognate subjects) – become a more valid method for aligning subject standards.

Percentage of marks on file at awarding: consequences for 'post-awarding drift' in cumulative grade distributions

Dhillon, D. (2004)

Awarding meetings are conducted with the aim of maintaining year-on-year, inter-specification, inter-subject and inter-awarding-body comparability in standards. To that end, both judgemental and technical evidence is implemented to facilitate grade boundary decisions. The difficulty arises when not all of the candidate mark data has been fully processed by the time of the award; hence, grade boundaries that appear to produce seemingly sensible grade distributions at award may change once all of the data has been re-run. Two methodologies were employed in an effort to investigate the degree of post-awarding drift that may occur in outcomes as a result of incomplete awarding data. First, empirical data from actual re-run GCE and GCSE awards during the summer 2003 series was collated and analysed. A large number of simulations were conducted in which different proportions of data were excluded from final GCE data sets according to two models designed to mimic the different kinds of late marks expected from the awarding databases.

Only a quarter (six out of twenty-four) of the re-run GCE awards demonstrated outcome changes of greater than one per cent at either key grade boundary. Post-awarding drift for the GCSEs was conspicuously more pronounced, especially at grade C, possibly due to the tiered nature of the specifications and/or the more heterogeneous nature of candidates and centres compared with GCE.

With respect to the simulations, although the overall magnitude of the changes between final and simulated outcomes varied according to subject, a consistent pattern was observed complying with the Law of Diminishing Returns. While increasing the percentage of candidates did decrease the

absolute difference between final and simulated outcomes, after a certain point this benefit became considerably less evident and eventually tended to tail off. While there are some limitations to the conclusions, both the empirical and simulated GCE data suggest that a lowering of the 'safe' cut-off point from 85- to 70-per cent fully processed at the time of GCE awards is unlikely to produce excessive changes to awarding outcomes that could compromise the approval of awards.

Inter-subject standards: An investigation into the level of agreement between qualitative and quantitative evidence in four apparently discrepant subjects

Jones, B. (2004)

The last two years have seen expressions of renewed concern, both in the press and by the QCA, about a perceived lack of comparability of standards between different subjects, particularly at GCE level. Research in this area has been relatively limited, largely because the caveats and assumptions that have to be made for both quantitative and qualitative approaches tend to undermine the validity of any outcomes. The methodological problems facing subject pairs analysis – one of the common statistical approaches – are rehearsed via a literature review.

A small research exercise investigated four subjects – two were deemed 'severe' and two 'lenient' by this method – that were identified by the press in 2003 for being misaligned. Putative grade boundaries that would bring these subjects into line with each other, according to the subject pairs definition, were calculated; scripts on these boundaries for the written units were pulled. The units' principal examiners were asked to identify where, on an extended grade scale, they thought the scripts were situated. The examiners for the 'severe' subjects, whose boundaries had been lowered, were quite accurate in placing the scripts; the examiners for the 'lenient' subjects, whose boundaries had been raised, were not only less accurate but tended to identify the scripts as low on the scale. The discussion considers why this might be the case, and whether the findings merit a more comprehensive investigation in view of the substantial political and practical problems.

Inter-subject standards: An insoluble problem?

Jones, B., Philips, D. and van Krieken R. (2005)

It is a prime responsibility of all awarding bodies to engender public confidence in the standards of the qualifications they endorse, so that they have not only usefulness but credibility. Although guaranteeing comparability of standards between consecutive years is relatively straightforward, doing so between different subjects within the same qualification and with the same grading scheme is a far more complex issue. Satisfying public and practitioner opinion about equivalence is not easy – whether standards are established judgementally or statistically or, as in most contexts, a mixture of the two. Common grade scales signify common achievement in diverse subjects, yet questions arise as to the meaning of that equivalence and how, if at all, it can it be demonstrated. With the increase in qualification and credit frameworks, diplomas and so forth, such questions become formalised through the equating of different subjects and qualifications – sometimes through a system of weightings.

This paper is based on two collaborative presentations made to the International Association for Educational Assessment (IAEA) conferences in 2003 and 2004. It summarises some recent concern about inter-subject standards in the English public examination system, and proceeds to describe three systems' use of similar statistical approaches to inform comparability of inter-subject standards. The methods are variants on the subject pairs technique, a critique of which is provided in the form of a review of some of the relevant literature. It then describes New Zealand's new standards-based National Qualifications Framework, in which statistical approaches to standard setting, in particular its pairs analysis method, have been disregarded in favour of a strict criterion-referenced approach. The paper concludes with a consideration of the implicit assumptions underpinning the definitions of inter-subject comparability based on these approaches.

Regulation and the qualifications market

Jones, B. (2011)

The paper is in four main sections. 'A theoretical framework from economics' introduces the conceptual framework of economics, in which the qualifications industry is seen as an operational market. Part 1 of this section describes the metaphors used to describe qualifications and their uses, and how these metaphors form, as well as reflect, how educational qualifications are perceived, understood and managed. Part 2 then summarises four typical market models as a background to understanding the market context

of the qualifications industry. Part 3 defines this context more closely, drawing particular attention to the external influences and constraints on it, from both the supply and demand sides. The following section ('Where have we been?') is a survey of general qualifications provision in England since the mid 19th century, which indicates how the industry has evolved through different types of market context, and, latterly, how statutory intervention and regulation has increased. 'Where are we now?' describes the 2009 Education Act and its implications and aftermath, particularly the significant changes to regulatory powers it introduced; and how, via various subsequent consultation exercises, it appears these changes are intended to be applied. Drawing on some of the information and issues raised explicitly or implicitly in the previous sections, the final section ('Where are we going? Regulation in a market context') considers the issues facing Ofqual following the 2009 Act, and looks to the possible direction, nature and implications of future regulatory practice.

Setting the grade standards in the first year of the new GCSEs

Pointer, W. (2014)

Reformed GCSEs in English, English Literature and Mathematics are being introduced for first teaching from September 2015, with the first examinations in summer 2017. Other subjects are being reformed to start the following year, with first examinations in summer 2018. The new specifications will be assessed linearly, and will have revised subject content and a numerical nine-point grade scale.

This paper looks at the results of simulations that were carried out to inform how the new grading scale for GCSEs will work. It discusses the pitfalls associated with various ways of implementing the new grade scale and highlights potential problems that could arise. It also evaluates the final decisions made by Ofqual. The paper focuses specifically on issues relating to the transition year, not subsequent years.

Ofqual has decided that the new grading scale should have three reference points: the A/B boundary will be statistically aligned to the 7/6 boundary; the C/D boundary will be mapped to the 4/3 boundary; and the G/U boundary will be mapped to the 1/U boundary. This will aid teachers in the transition to the new grading scale, and will also aid employers and further education establishments to make more meaningful comparisons between candidates from different years. If possible, pre-results statistical screening will be used to ensure comparability between awarding organisations at all grades, not just those that have been statistically aligned, by means of predictions based on mean GCSE outcomes.

Aggregation, grading and awarding

Aggregation, grading and awarding are critical processes in the examination cycle. Once the examination has been marked, the marks from individual questions are summed to give a total for each examination paper; the paper marks are then added together to give a total for the examination as a whole. This process is termed aggregation. The total examination scores are then converted into grades via the process of awarding – this determines the outcome for each student in terms of a grade that represents an overall level of performance in each specification. In ongoing examinations, the aim is to maintain standards in each subject both within and between awarding organisations – and across specifications – from year to year.

The process of mark aggregation is affected by various factors such as the nature of the mark scales and the extent to which each individual component influences the overall results. Ensuring that candidates who are assessed on different occasions are rewarded equally for comparable performances has been a key issue in recent years, and relates to modular (or unitised) examinations. Candidates certificating on any given occasion will have been assessed on each unit on one of several different occasions, and may have retaken units. Aggregation and awarding methods must place the marks obtained for any particular occasion onto a common scale, so that these marks can then be aggregated fairly across the units.

As new examination papers are set in every specification each time the examination is offered, a new pass mark (or grade boundary) has to be set for each grade. Apart from coursework (which follows the same assessment criteria year on year and therefore, generally speaking, the grade boundaries are carried forward in successive years), grade boundaries cannot be carried forward from one year to the next because the papers vary in difficulty and the mark schemes may have worked differently, with the result that candidates may have found it easier or more difficult to score marks. To ensure that the standards of attainment demanded for any particular grade are comparable between years, the change in difficulty has to be allowed for.

Awarding meetings are held to determine the position of the grade boundaries. In these meetings, a committee of senior examiners compare candidates' work from the current year with work archived from the previous year, and also

review it in relation to any published descriptors of the required attainment at particular grades. Their qualitative judgements are combined with statistical evidence to arrive at final recommendations for the new grade boundaries.

Essentially, the role of each awarding committee is to determine (for each examination paper) the grade boundary marks that carry forward the standard of work from the previous year's examination, or that set standards in an entirely new examination. In the latter scenario, this has recently involved carrying forward standards from the previous legacy specification; however, there will be forthcoming challenges in setting standards for new specifications that have no previous equivalent.

Much work has been carried out to investigate various aspects of the awarding process – including the nature of awarders' judgements and the way in which scrutiny should be carried out – and in developing new statistical approaches as a (generally) more reliable tool for maintaining standards, some of which are summarised below.

A general approach to grading

Adams, R. M. and Mears, R. J. (1980)

This paper outlines the theory of a general approach to the grading of examinations. It points out that, for a two-paper examination, Ordinary Cartesian axes in the plane can be used to represent the paper one and paper two scores. Because each paper has a maximum possible score, and because negative scores cannot occur, attention can be restricted to a rectangular region in the first quadrant. Further, because marks are only awarded in whole numbers, candidates will only occur in this rectangular space at points (x, y) where x and y are integers. Thus, the score space can be represented as a rectangular array of points in the first quadrant. The paper goes on to consider the representation of a variety of grading schemes in the score space, including the use of component grade hurdles and schemes for limiting the nature of compensation between the components.

Norm and criterion referencing of performance levels in tests of educational attainment

Cresswell, M. J. and Houston, J. G. (1983)

This paper considers a basic test of educational attainment: a spelling test in which the candidates have to spell 100 words correctly, all words being equally creditable. Two performance levels are defined: 'pass' and 'fail'.

The nature of norm and criterion referencing is discussed using this simple example. Findings indicate that it is difficult to specify performance criteria, even for a unidimensional test for which two performance levels are needed – only one of which has to be defined since the second is residual. It is then argued that when tests of educational attainment in school subjects are brought into the discussion, the difficulties are greatly multiplied. The complex matrix of skills and areas of knowledge implied by what is being tested means that there will be many different routes to any given aggregate mark. In following these routes, candidates will have satisfied different criteria. It will be impossible to find a common description that in any way adequately describes all the routes leading to that given aggregate mark. The specification of subject-related criteria is a daunting task: if only a few crucial criteria are specified, many candidates who satisfy them may seem to fail to satisfy more general but relevant ones. On the other hand, if very complex multi-faceted criteria are specified, few candidates will succeed in meeting them fully.

Profile reporting of examination components: how many grades should be used?

Cresswell, M. J. (1985)

This paper considers the case in which component grades are reported for each candidate. It discusses the existence of apparent anomalies between the component grades and the grades for the examination as a whole – if the latter are awarded on the basis of candidates' total scores. The paper shows that, if the whole examination is reported in terms of the GCSE grade scale, then the total incidence of such anomalies is minimised by the use of a scale of three or four grades for the components. However, two types of apparent anomalies are identified. The more problematic ones occur less frequently as the number of component grades is increased. The paper recommends the use of an eight-point scale for any component grades reported for GCSE examinations.

Examination grades: how many should there be?

(*British Educational Research Journal*, Vol. 12, No. 1)

Cresswell, M. J. (1986)

There is no generally accepted rationale for deciding the number of grades that should be used to report examination results. Two schools of thought

on this matter have been identified in the literature. One view is that the number of grades should reflect the reliability of the underlying mark scale. The other view focuses upon the loss of information incurred when the mark scale is reduced to a number of fairly coarse categories. The first of these views usually implies the adoption of a relatively small number of grades; the second view implies the use of a considerably larger number of grades. In this paper, the various factors that determine the relative merits of these two schools of thought are considered in relation to the different functions which examinations fulfil.

Placing candidates who take differentiated papers on a common grade scale

(*Educational Research*, Vol. 30, No. 3)
Good, F. J. and Cresswell, M. J. (1988)

Three methods of transferring marks from differentiated examinations on to a common grade scale are compared. Equi-percentile scaling and linear scaling prior to grading gave very similar grades. However, grading the different versions of the examination separately – without scaling the component marks for difficulty – resulted in the award of different grades to a substantial proportion of candidates. The advantages and shortcomings of each method are considered and also whether a scaling method or separate grading is to be preferred. It is concluded that a scaling method should be used, and that the grades from linear scaling are likely to be the most satisfactory.

Combining grades from different assessments: how reliable is the result?

(*Educational Review*, Vol. 40, No. 3)
Cresswell, M. J. (1988)

Assessment usually involves combining results from a number of components. This has traditionally been done by adding marks and the issues raised are discussed in most books on assessment. Increasingly, however, there is a need to consider ways of providing an overall assessment by combining grades from component assessments. This approach has been little discussed in the literature. One feature of it, the likelihood that the overall assessment will be less reliable than one based upon the addition of marks, is explored in depth in this paper. The reliability of the overall assessment is shown, other things being equal, to depend upon the number of grades used to report achievement on the components.

It is concluded that the overall assessment will be satisfactorily reliable only if the number of grades used to report component achievements is equal to, or preferably greater than, the number used to report overall achievement.

Fixing examination grade boundaries when components scores are imperfectly correlated

Good, F. J. (1988)

This paper considers two methods of combining component grade boundaries. Using one method, the component grade boundaries are added to give the corresponding examination boundaries. This procedure is called the Addition Method. The other method finds the mean percentage of candidates, weighted if appropriate, that reach each component boundary and defines each corresponding examination boundary as the mark that cuts off the same percentage of candidates on the examination score distribution. This is called the Percentage Method. The methods are considered in terms of the assumptions that are required for each, and the extent to which these assumptions are realistic. The effects of three factors on the position of the grade boundaries fixed by the Percentage Method are also considered. These factors are differing proportions of candidates reaching the boundaries on different components, differing component standard deviations, and the application of different component weights.

Grading the GCSE

(The Secondary Examinations Council, London)
Good, F. J. and Cresswell, M. J. (1988)

In some GCSE examinations, candidates at different levels of achievement take different combinations of papers. The papers taken by candidates who aspire to the highest grades are intended to be more difficult than those taken by less accomplished candidates. The main aim of the Novel Examinations at 16+ Research Project was to investigate the issues that arise when grades are awarded to candidates who have taken an examination of this type; that is, an examination involving differentiated papers. The fundamental problem with which the project was concerned was that of making fair comparisons between the performances of candidates who have taken different papers that are set at different levels of difficulty and cover different aspects of the subject being examined. The ability of awarders to give candidates grades that are fair in this sense was investigated. Methods by which marks achieved on different versions of an examination can be adjusted so as to lie on a common scale were also

studied. The alternative to differentiated papers – common papers that are taken by every candidate – was also briefly considered as a means of providing differentiated assessment.

Grading problems are minimised by the use of common papers; the main difficulties lie in producing papers that reward all candidates' achievement appropriately. One of the approved methods of doing this – the placing of questions (or part questions) on an incline of difficulty – was found not be theoretically viable and it is also difficult to achieve in practice. The other commonly proposed technique of differentiated assessment in common papers is the use of questions that are neutral in difficulty and can be answered at a number of distinct levels of achievement. However, there must be doubt as to whether candidates taking such questions always respond at the highest level of which they are capable.

For the purpose of grading differentiated papers, it is suggested that grades can be defined as comparable if they are reached by the same proportion of a given group of candidates. However, this definition was not consistent with the grade awarders' judgements of comparable performances. The awarders tended to consider fewer candidates to be worthy of any given grade on harder papers or, alternatively, that more candidates reached the required standards on easier papers. While there may be circumstances in which too strict an adherence to statistical comparability (as defined above) would be incorrect, grading should be done using a method that guides the awarders towards judgements that are statistically consistent within an examination. Unless this guidance is given, any particular grade tends to be more easily achieved from the easier version of a differentiated papers examination. That is, candidates who enter for a harder version tend to get lower grades than they would have got if they had entered for an easier version. This effect was shown clearly in this study, in which some candidates took the papers for two versions of the experimental examinations.

The study covered various methods of grading candidates in terms of a common grade scale when they have taken different combinations of papers. In general, methods involving adding together candidates' marks from the papers and then fixing grade boundaries on the scale of total marks were superior to methods that involved grading each paper and then combining the candidates' paper grades into an overall grade. It was concluded that, where an examination involves candidates taking one of two alternative versions with only part common to all candidates, the paper marks should be transferred to a common mark scale (using conventional scaling techniques) before they are added and the examination graded as a whole.

Finally, where the harder version of an examination comprises all the papers from the easier version together with an optional extension paper, candidates entered for the harder version should also be graded as if they

had been entered for the easier version and should then be awarded the better of their two grades. Further, it is desirable for the extension paper (taken by the more able candidates for the award of higher grades) to be given at least as much weight as the combination of easy version papers. If this is not done, the harder version may not discriminate adequately between the most able candidates.

The discrete nature of mark distributions
Delap, M. R. (1992)

In 1992, new procedures were implemented at award meetings. Awarding committees were asked to write a rationale for any recommendation that suggested a change in the cumulative proportion of candidates obtaining grades A, B and E of more than one, two and three per cent respectively. Many awarders felt that the statistical limits were too severe. This paper discusses effects that are caused by the discrete nature of mark distributions. The method used to compute the statistical limits of one, two and three per cent required the assumption that the mark distributions were continuous. The paper shows that this is not necessarily an appropriate assumption. A new method of computing statistical limits is presented that takes account of the discrete nature of the mark distribution.

Aggregating module tests for GCSE at KS4: choosing a scaling method
Cresswell, M. J. (1992)

In modular GCSE examinations, candidates who have taken different sets of module tests must all be awarded comparable grades on the basis of the combination of all their module assessments and their terminal examination assessment. However, module tests from the different tiers are deliberately made to differ in difficulty. Therefore, it is not possible to simply add up each candidate's total score from all the module tests that he or she has taken, since the result will vary depending upon the tiers of those tests. It is necessary to render the scores from each module test comparable by some scaling process before candidates' total module scores are computed and added to their corresponding terminal examination scores. This paper outlines some of the methods of doing the required scaling and indicates the conditions under which each may be used.

Aggregation and awarding methods for national curriculum assessments in England and Wales: a comparison of approaches proposed for Key Stages 3 and 4

(*Assessment in Education*, Vol. 1, No. 1)
Cresswell, M. J. (1994)

Most educational assessment involves aggregating a large number of observations to form a smaller number of indicators (for example, by adding up the marks from a number of questions). The term 'awarding' refers to any subsequent process for converting aggregated raw scores onto a scale that facilitates general interpretations. This paper explores some of the theoretical and practical issues involved in aggregation and awarding by considering the relative merits of two methods: the method used at the end of National Curriculum Key Stage 3 in 1993 and a more conventional method proposed for assessment at the end of Key Stage 4. It is shown that aggregation and awarding procedures like those used in 1993 at Key Stage 3 are unlikely to produce results that are as fit for the common purposes of assessment as more conventional procedures.

'Judge not, that ye be not judged'. Some findings from the Grading Processes Project

Paper given at an AEB research seminar on 21 November 1997 at Regent's College, London
Cresswell, M. J. (1997)

This is one of the main reports from a seven-year investigation into awarding. It concentrates on the empirical work of the project and describes the findings of an observational study of conventional examination awarding meetings that aimed to provide a full description and better understanding of the way in which judgement operates within the awarding process. In particular, the evidence that is actually used by awarders as a basis for their judgements is described and so are the ways in which they use that evidence.

The study concluded that examination standards are social constructs created by groups of judges, known as awarders, who are empowered, through the examining boards as government-regulated social institutions, to evaluate the quality of students' attainment on behalf of society as a whole. As a result, standards can be defined only in terms of human evaluative judgements and must be set initially on the basis of such judgements.

The process by which awarders judge candidates' work is one in which direct and immediate evaluations are formed and revised as the awarder reads through the work. At the conscious level, it is not a computational process and it cannot, therefore, be mechanised by the use of high-level rules and explicit criteria.

Awarders' judgements of candidates' work are consistently biased because they take insufficient account of the difficulty of examination papers. Such judgements are therefore inadequate, by themselves, as a basis for maintaining comparable standards in successive examinations on the same syllabus. The reasons for this are related both to the social psychology of awarding meetings and to the fundamental nature of awarders' judgements.

The use of statistical data alongside awarders' judgements greatly improves the maintenance of standards, and research should be carried out into the feasibility of using solely statistical approaches to maintain standards in successive examinations on the same syllabus. A broadening of the range of interest groups explicitly represented among judges, who initially set standards on new syllabuses should also be considered.

Can examination grade awarding be objective and fair at the same time? Another shot at the notion of objective standards

Cresswell, M. J. (1997)

This paper contests the notion that examination standards are, or can be made into, objective entities (some variety of Platonic form, presumably) that sufficiently skilled judges can recognise using objective procedures. Unease about the subjective nature of examination standards is misplaced, and any attempt to make awarding fairer by the objective use of explicit criteria and aggregation rules is fundamentally misconceived. This approach is not, necessarily, fair at all and is based upon a conception of judgement that is highly questionable. The paper proposes an alternative model for the process of evaluation that is consistent with a modern understanding of the nature of critical analysis. This model is compatible with the recognition that examination standards are not objective but are social constructs created by groups of judges, known as awarders, who are empowered, through the examining boards as government-regulated social institutions, to evaluate the quality of students' attainment on behalf of society as a whole.

The effects of consistency of performance on A-level examiners' judgements of standards

(*British Educational Research Journal*, Vol. 26, No. 3)

Scharaschkin, A. and Baird, J. (2000)

One source of evidence used for the setting of minimum marks required to obtain grades in General Certificate of Education (GCE) examinations is the expert judgement of examiners. The effect of consistency of candidates' performance across questions within an examination paper upon examiners'

judgements of grade-worthiness was investigated, for A-level examinations in two subjects. After controlling for mark and individual examiner differences, significant effects of consistency were found. The pattern of results differed in the two subjects. In Biology, inconsistent performance produced lower judgements of grade-worthiness than consistent or average performance. In Sociology, very consistent performance was preferred over average consistency. The results of this study showed that a feature of the examination performance that was not part of the marking scheme affected grading decisions. It is concluded that examiners' judgements of standards should be supported by other sources of evidence, such as statistics.

Is the whole worth more than the sum of the parts? Studies of examiners' grading of individual papers and candidates' whole A-level examination performances

(*Educational Studies* 28, 2)
Baird, J. and Scharaschkin, A. (2002)

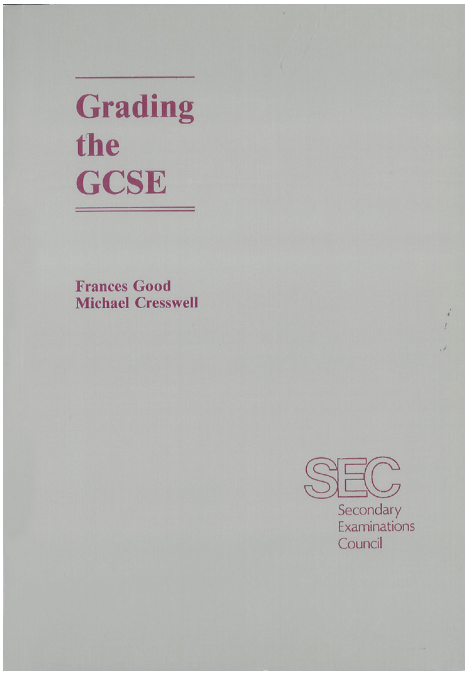
Typically, students are assessed on elements of their performance, and it is assumed that the sum of marks for these elements will be just as impressive as the students' whole performances. Examiners might expect more for a particular grade if they only see parts of the students' work separately. Two experiments were carried out comparing examiners' judgements of the grade-worthiness of candidates' A-level examination work at question-paper level and at subject level. The results of both studies suggested that examiners may have compensated for the different aspects of the subject tested in different question papers when they made holistic judgements, but did not make this compensation when they made question-paper judgements. Tunnel vision effects are likely to be greater in the AS/A2 examinations than those found here, because the examinations will be broken into smaller parts.

Awarding objective test papers: is there a correct answer?

Meyer, L. (2008)

In AQA, Objective Tests (OTs) in GCSEs have, historically, been awarded via a process known as 'back calculation' from other components. More recently, other methods – such as Angoff (GCE) and KS3 predicted outcomes/IRT (GCSE Science) – have also been employed. Back calculation

and the use of predicted outcomes do not involve professional judgement, whereas the Angoff approach does. However, the latter method has had mixed success when used in the current AQA GCEs. In 2004 and 2005, trials of awarding OTs using the Bookmark method were carried out (Fowles, 2004 & 2005), but these did not give promising results and the approach was not pursued. This paper reviews the most recent literature on the principal current methods of awarding that can readily be applied to OTs, and which incorporate professional judgement into the process. The methods are compared, and the apparent pros and cons of each approach summarised, to provide an overall evaluation of the current scenario. The paper is also intended to serve as a basis for discussion as to how to conduct the award of the new AQA Diploma OT unit in January 2009, and the awards of other new OT units in the future.



The Good and Cresswell effect: F. J. Good and M. J. Cresswell's seminal book *Grading the GCSE*, published in 1988, had a deep impact on awarding procedures

It's a long, long time from November to June

Wheadon, C. (2009)

Since 1918, if not before, the maintenance of standards over time in the English examination system has used approaches that assume that large cohorts of candidates sit their examinations at the same time of year, every

year, after following a similar programme of study over a similar time period. These assumptions present barriers to the modernisation of the examination system. Firstly, the personalisation policy agenda seeks to deliver a personalised classroom with a personalised examination timetable by 2020. Secondly, the delivery of on-screen assessment is currently being hampered by the limitations on the number of candidates that can be tested on-screen in any centre in any one sitting. Multiple parallel versions of tests would allow longer testing windows, but would pose the standard-setting problem of multiple heterogeneous populations.

Item Response Theory (IRT) test-equating approaches would seem to hold the answer as the parameters that characterise an item do not depend on the ability distribution that characterises the examinees. However, IRT approaches depend on strong statistical assumptions that do not hold precisely in real testing situations. This research was undertaken to investigate the extent to which the invariance of item parameters would hold for a post-equating non-equivalent group design intended to maintain standards between a June and a November test session.

Assuring comparability of scores across different test versions

Wheadon, C. (2011)

One of the major barriers to delivering high-stakes on-screen examinations in the UK is the need for all candidates to be examined at the same time. This constraint exerts a substantial logistical strain on technology resources in schools. If candidates could be tested at different times, using different test versions the strain on resources could be alleviated. For such an approach to work, suitable systems must be in place to ensure that standards can be maintained across the test versions.

To examine the robustness of various approaches to this problem, a trial was undertaken using four versions of a French reading test with over 10,000 15 and 16 year olds. The test versions contained common items and were randomly allocated to candidates who logged onto the test delivery system. Prior achievement measures were also gathered on all candidates. The design of the experiment meant that three equating designs could be compared; Equivalent Groups (EP), Item Response Theory (IRT), and an underlying linked construct of general ability. The results revealed very little discrepancy between the methods. The success of the equating based on the underlying linked construct suggests that this method, rather than the repetition of items between test versions (which is a security risk) could supplement random allocation of test versions to ensure robust standards between test versions.

Non-examination assessment

Non-examination assessment is a term recently coined by Ofqual to describe any type of assessment that is not a timed written examination. Previous terms have included internal assessment, coursework and controlled assessment; unfortunately, all of these have slightly different meanings. For example, coursework may not always be internally assessed, and non-examination assessment encompasses practical tests with a visiting examiner as well as coursework. Controlled assessment was a definition introduced by Ofqual for the GCSE specifications that were first certificated in 2011, and it signified a greater degree of rigour and supervision than had customarily been associated with coursework.

Coursework was introduced on a large scale in the first GCSE examinations in 1988. At that time, there was a regulatory requirement for each syllabus to include *at least* the amount of coursework specified for the subject in question. However, even by the early 1990s, concerns were raised about the reliability of coursework marks.

The papers in this section give a flavour of some of the research that has been conducted over the past four decades.

Differences in marks awarded as a result of moderation: some findings from a teacher-assessed oral examination in French

(*Educational Review*, Vol. 40, No. 3)
Good, F. J. (1988)

It is unlikely that all teachers will apply identical standards and criteria when marking school-based components of examinations, so moderation procedures are required to check and, if necessary, adjust their marks. This paper considers differences between the marks awarded by teachers and moderators in a French oral examination when the moderators re-marked the work without knowledge of the teachers' marks. It finds that although teachers were generally more lenient than moderators, their rank orders of the candidates were similar. A general statistical method of transforming the marks to a common mark scale, when there is little overlap between the assessment objectives that are tested in the internal and external components, is outlined. The implications of the differences in the judgements of teachers and moderators for the statistical procedures are considered. It is suggested that a structural regression method should generally be used to transform the marks, and that all candidates should be awarded the teacher's mark adjusted in this way – rather than the moderator's mark or the raw teacher's mark.

A method of moderation of school-based assessments: some statistical considerations

(*The Statistician* 37)
Good, F. J. (1988)

Most GCSE examinations include a school-based component that is marked by the teacher. Adjustments to these marks may be required to give a common measurement scale for all centres. Variations on a statistical method of transforming the marks are considered, together with the assumptions that are inherent in each. Formulae by which the accuracy of the resulting marks can be assessed are given. The meaning of the adjusted mark and the implications of using the moderator's or teacher's mark instead are considered. A small data set is analysed to illustrate the suggested procedures and the differences in the marks awarded with each. It is concluded that a structural regression line should be used to calculate the adjusted marks, and that these marks should generally be used in preference to either the teacher's or moderator's marks.

Reliability of judgements made by coursework moderators: synopsis

Taylor, M. (1991)

A sample of moderated coursework from the summer 1991 examinations has been collected in three subjects (English Syllabus A, Mathematics Syllabus A and History Syllabus 1) at GCSE and in one subject (Psychology) at A-level. In each subject, two moderators will be asked to re-moderate a number of pieces of work, and to report on aspects of the moderation procedure.

The main aim of the research is to investigate the extent of the agreement or disagreement between the marks awarded by different moderators – including the original moderator. It is hoped to identify those areas where divergence is most likely to occur, and to try to ascertain the factors that may cause moderators to differ in their judgements.

Teacher moderation systems

Report for the National Assessment Agency

Taylor, M. (2005)

In most current GCE, VCE, GCSE and GNVQ specifications, coursework comprises one or more defined tasks (either set by the awarding body or based on criteria defined by the awarding body) and is marked by centres. Because the assessment is of an end-product, the work of a sample of candidates from each centre can be re-marked by a moderator, and adjustments to the centre's marks of all candidates at that centre can be determined, where necessary, based on a comparison of the centre marks and moderator marks for the candidates in the sample. This is known as moderation by inspection.

Although major changes to coursework are not now expected, the Working Group on 14-19 Reform (2004) had favoured a move away from set-piece coursework tasks to a more open-ended style of teacher assessment, possibly based on the general work of candidates during the course. One of the consequences of such a move would have been to remove the possibility of moderation in the manner described above. Therefore, thoughts turned to the use of statistical methods as part of the monitoring process, in order to check whether centres appear to be marking to the correct standard.

The present study has two parts. The first part investigates the use of one or more externally assessed components to moderate an internally assessed coursework component, using statistical methods. The second part uses centres' estimated grades as a proxy for teacher assessment

and investigates the effect on candidates' overall results of replacing operational marks with these 'teacher assessments', either moderated or unmoderated.

The effect of video moderation on moderator marks
Spalding, V. and Joyner, S. (2013)

Visiting moderation is currently used for performance-based controlled assessment units. This is a costly and inefficient process. Video moderation has been proposed as an alternative procedure. This paper investigates the effect of video moderation on moderator marks via a post-hoc mark-comparison exercise. The results show that video moderation marking was more severe than the original visiting moderation. An ANCOVA analysis revealed that the marking was fairly inconsistent, though this study cannot determine whether video moderation is any more or less reliable than visiting moderation. The paper concludes that video moderation is feasible. However, it may result in more severe marking and would require a large-scale improvement in the provision of video evidence from centres.

Quality of marking

Millions of high-stakes written examination papers are marked every year; ensuring that the marks are accurate is of critical importance. High-quality marking relies on a robust underlying curriculum, strong marker training and rigorous testing (see 'Assessment design', pp. 60–65).

Some assessment systems are more straightforward than others. For example, the SAT and ACT tests used for college admission in the US rely on multiple-choice questions; each question is either right or wrong. Marking is simply a matter of checking that each score has been accurately recorded and the total marks tally up. There is no need to apply professional judgement – the marking is entirely objective – and so marking reliability is extremely high. This type of examination can be a useful way of testing certain levels of understanding.

However, in the English examination system, we assess students' abilities in a variety of ways. We ask them to write essays, draw graphs, perform music and construct arguments. This ensures validity of measurement, but means that marking is more complex.

CERP – like its predecessors – provides research evidence to inform question paper and mark scheme design. We have investigated the features of mark schemes that are more likely to lead to reliable marking, and identified the types of questions that are difficult to mark. Our researchers have also analysed the characteristics of a good marker and the importance that standardisation plays in improving reliability. Latterly, attention has turned to electronic marking and innovative ways of assessing quality via comparative judgements.

The following abstracts highlight how this process has evolved over the past four decades.

Reliability of marking in eight GCE examinations

(*British Journal of Educational Psychology*, Vol. 48)
Murphy, R. J. L. (1978)

Eight GCE examinations that contained mainly free-response questions were investigated in terms of their marking reliability. The scripts of 200 randomly selected candidates from each subject were re-marked by a senior GCE examiner, and these marks were compared with the marks awarded previously as a result of team-marking procedures. These comparisons revealed differences between the reliability of marking of examinations in different subject areas, and also between different papers within individual examinations. The results are discussed in terms of differences between examinations in different subject areas, the effect of increasing the number of questions in an examination, and the effect of including questions other than free-response questions. All of these factors have an effect on the reliability of the marking of these examinations. The results are also compared with other estimates for similar examinations.

Post-results re-marks: how should they be used?

Cresswell, M. J. (1996)

When a second marker re-marks a candidate's work post results, a second assessment is obtained to stand alongside the original one. The two assessments will rarely agree completely and the question arises of how to award a grade to the candidate that takes account of both assessments in an appropriate way. For example, if the second assessment was independent of the first, the two were equally reliable and neither was biased in terms of severity or lenience, then the mean of the two marks would give the best measurement available. However, under other conditions, other approaches may be more appropriate. Three different approaches are considered in this paper; it is concluded that the current practice of grading the candidate on the basis of the re-mark and issuing the new grade if it is higher than the original grade – no matter how small the mark change involved – is the best available approach.

Marking consistency over time

(*Research in Education*, 67)
Pinot de Moira, A., Massey, C., Baird, J., and Morrissy, M. (2002)

It is of great importance that examination bodies award candidates grades that correctly reflect the quality of work presented, because examination results can affect life chances. Marking should be at a common high standard and free from bias, otherwise some candidates are placed at an unfair advantage and others at an unfair disadvantage. In the past, there has been much research into the reliability of marking (for example, Newton, 1996; Branthwaite et al., 1981; Murphy, 1978) and into the existence of marking bias (for example, Baird, 1998; Massey, 1983). Few studies have considered the longitudinal reliability of marking: determining the influence that script sequence, or moment in time, can exert on the accuracy of assessment. Spear (1997) examined the biasing influence of contrast effects and found that if a good piece of work was assessed before work of a lower standard, then the poorer-quality work would be assessed more harshly. On the other hand, if higher-standard work was preceded by a piece of lesser quality it would be assessed more favourably. In a ten-year retrospective study, Lunz and O'Neill (1997) showed that, although individual judges

vary in their level of leniency, the leniency of most judges remains internally consistent throughout, in spite of retraining. This study corroborated earlier evidence presented by Lunz and Stahl (1990), where it was shown – albeit over a substantially smaller timescale – that judges' leniency is reasonably consistent over time, notwithstanding some variations across grading period.

Such findings are of considerable interest in the context of the public system of examining in England, Wales and Northern Ireland. After standardisation meetings, examiners assessing any of the national qualifications are required to mark an allocation of scripts over a period of two to three weeks. Anecdotal evidence provided by senior examiners and awarding body staff suggests that, as time since standardisation increases and pressure to complete the marking exercise increases, so the accuracy of examiners' marking declines. However, as the pressure increases for examiners, so does the pressure for those monitoring examiner performance. Workload is prioritised to focus on those examiners causing the most concern and, as such, the anecdotal evidence may be flawed, as the checks may be biased towards the poorest examiners at the end of the marking period. This study sets out to explore the view that marking accuracy decreases through the marking period, in the context of A-level examinations.

Evaluation of an e-marking pilot in GCE Chemistry: effects on marking and examiners' views

Fowles, D. (2002)

GCE A-level Chemistry examiners who participated in an NCS Pearson/AQA e-marking pilot in April 2002 provided research data from two sources. The first source was questionnaire responses and other comments volunteered by the examiners in writing or orally. These have been brought together in the first section of this report, which focuses on the examiners' attitudes to e-marking. The second source of data was the examiners' electronically captured marks, matched to the original – 'conventional' – marks awarded in the January 2002 GCE Chemistry CHM1 unit examination from which the scripts for the e-marking pilot were drawn. These two versions of the marking are compared in the second section of this report.

Examiner background and the effect on marking reliability

Pinot de Moira, A. (2003)

This report discusses a study of the background of examiners and the marks they give. It arises from recommendations of the independent panel report on maintaining GCE A-level standards (Baker, McGraw and Lord Sutherland

of Houndwood, January 2002). Even though there is little published literature that relates reliability to examiner characteristics, the presented work is set in the context of existing marking reliability research.

Data from a sample of 21 AQA A2 units, marked by 356 examiners in summer 2002 has been analysed by fitting four multilevel models. Each model considers a different aspect of marking reliability, as represented by four statistical measures: the difference between senior examiner and assistant examiner mark; the absolute difference between senior examiner and assistant examiner mark; the probability of a numerical adjustment having been made to the assistant examiner's marks; and the examiner performance rating. Unit, examiner, centre and candidate level independent variables are included where they explain a significant amount of variation in the dependent variable.

The study identifies no link between personal characteristics and marking reliability. Evidence suggests that reliability is more closely related to features of an examiner's allocation and the idiosyncrasies of individual subjects. The models produce some evidence to support the argument that the work of more able candidates is harder to mark, as is the work of candidates from independent and selective establishments.

Questionnaire responses from principal examiners shed some light on possible reasons for the observed centre-type differences. Recommendations are made for future research in the area, with a view to gaining a greater understanding of the influences on marking reliability, and to using this understanding to operational advantage.

Electronic marking with ETS software

Royal-Dawson, L. (2003)

Seven markers – three mathematics and four English – took part in an e-marking study using Educational Testing Service (ETS) software called the On-line Scoring Network (OSN). Markers' attitudes to the system and e-marking in general were recorded. The marks they awarded during the e-marking exercise, and those awarded conventionally afterwards, were collated and analysed.

Even though two of the markers (one English and one mathematics) do not use computers, all adapted well to the system, and were positive about it: the mathematics markers more so than the English ones. The English markers had reservations about the system's suitability for items assessing higher-order skills, such as writing, and the lack of access to other parts of a candidate's work to help corroborate their decisions. All markers prefer to mark at home, working the hours that suit them. Three of the markers did

not have computing facilities at home; these markers felt they would need to be more proficient at computing to use the system independently.

Twelve per cent of the items were e-marked by two markers as a means of monitoring the quality of marking. For English, there was less agreement for items with higher mark allocations. For mathematics, there was 98 per cent agreement. The markers were also required to mark 100 scripts conventionally at home. Per candidate, the mean total mark from conventional marking was calculated and compared to the total mark awarded through the OSN system. The size and direction of the differences between the two totals were not the same for the two subjects. For English, conventional marking tended to yield higher marks, and for mathematics, the differences between conventional marking and e-marking were higher and lower in equal measure.

The study did not explore the full power of the ongoing standardisation facility within the OSN software. Future pilots should investigate it to assist with the development of an appropriate model for the standardisation of e-markers. This model should include recommendations on the mark tolerance per item that is applied to accept or reject a marker's decision compared to a standard.

What makes marking reliable? Experiments with UK examinations

(*Assessment in Education: Principles, Policy & Practice*, Vol. 11, No. 3)
Baird, J., Greatorex, J. and Bell, J. (2004)

Marking reliability is purported to be produced by having an effective community of practice. No experimental research has been identified that attempts to verify, empirically, the aspects of a community of practice that have been observed to produce marking reliability. This research outlines what that community of practice might entail, and presents two experimental studies on the effects of particular aspects of a community of practice on examiners' marking reliability. In the first study, the impact of exemplar work is investigated: examiners were provided with mark schemes, and some examiners were provided with exemplar scripts and given feedback about the marking of those scripts. The second study explores the effects of discussion of the mark scheme: all examiners received mark schemes and exemplar scripts, but some examiners did not attend a coordination meeting. Neither procedure (use of exemplar scripts or discussion between examiners) demonstrated an improvement in marking reliability, which called into question the predictive utility of the theory of community of practice.

An investigation of targeted double marking for GCSE and GCE

(Part of a series of investigations by the National Assessment Agency – published by QCA)
Fearnley, A. (2005)

The aim of the current study was to investigate whether a system of double marking could be devised that would improve reliability without requiring more extensive marking by principal examiners. The study used two examination components; some of the examiners had marked without seeing each others' marks and annotations, and some had marked while seeing others' judgements.

This research considered whether double marking – using annotated or cleaned scripts – improved the reliability of marking, and, if so, by how much. Further, how much improvement in reliability is there when some markers do see other examiners' annotations, and when some markers do not see the evidence of any first-instance marking? A random allocation of examiners into pairs produced one method of pairing examiners for double-marking purposes. A second method was that of targeted pairings based on examiners' previous examiner performances. The effectiveness of each method was tested in the study.

Is teaching experience necessary for reliable marking?

Royal-Dawson, L. and Baird, J. (2006)

Although hundreds of thousands of markers are recruited internationally to mark examinations, little research has been conducted on the selection criteria that should be used. Many countries insist that markers have teaching experience, and this has frequently become embedded in the cultural expectations surrounding the tests. Shortages in markers for some of the UK's national examinations has led to non-teachers being hired to mark certain items, and changes in technology have fostered this approach. This study investigated whether teaching experience is a necessary selection criterion for a national curriculum English examination taken at age 14. Fifty-seven markers with different backgrounds were trained in the normal manner and marked the same 98 students' work. By comparing the marking quality of graduates, teacher trainees, teachers and experienced markers, this study shows that teaching experience was not necessary for most of the examination questions. Small differences in inter-rater reliability estimates on the Shakespeare reading and writing tasks were found, for example non-teachers were less reliable than those with teaching experience. A model for the selection of markers to mark different question types is proposed.

Features of a levels-based mark scheme and their effect on marking reliability

Pinot de Moira, A. (2013)

Levels-based mark schemes are commonly used in the marking of extended response items but, between specifications, there is little commonality in their design, formulation and application. This study establishes a list of the variable design characteristics between levels-based mark schemes and analyses marking reliability with reference to these characteristics. It finds that most of the variation in marking reliability is due to the vagaries of individual responses, which a holistic approach to item design might mitigate. It also recommends a number of small adjustments to mark scheme design that might improve marking reliability and increase the transferability of skills between the marking of different items, units and specifications.

Gains in marking reliability from item-level marking: is the sum of the parts better than the whole?

(*Educational Research and Evaluation: An International Journal on Theory and Practice*, Vol. 19, No. 8)

Wheadon, C. and Pinot de Moira, A. (2013)

Marking of high-stakes examinations in England has traditionally been administered by schools and colleges sending their examination papers directly to examiners. As a consequence, the work of one candidate has, historically, been marked by one examiner – as has work of an entire centre. Previous studies have suggested that the marking of both whole scripts and whole centres is liable to bias, caused either by examiner characteristics or the characteristics of the allocation of marking. This study used operational data from two geography papers that had moved from whole-script, whole-centre marking to item-level marking, and then back again, to quantify the gains in reliability from item-level marking. It found that there were substantial gains in the reliability of marking for the highest performing candidates when the marking was at item level. The reasons for these gains did not appear to be associated with any characteristics of a centre entry.

Assessment validity

In the 40 years that the AQA Research Committee has been in existence, the concept of test validity has changed dramatically, as has its status within assessment research generally. In 1975, influential theorist Samuel Messick tried to unify the disparate models of validity and argued that all validity was essentially construct validity – a perspective that came to dominate late 20th-century thinking on validity.

However, since the turn of the century, discourse has fragmented and several ongoing debates have emerged. One set of debates, for example, concerns the scope of validity, and focuses on the position of consequences in relation to validity and validation. Another set of debates focuses on the relationship between validity theory and its operability on a practical level in real assessment situations. There is a move to streamline the theory in recognition of the enormous amount of evidence required to comprehensively validate an assessment.

There is no question that the notion of validity is central to assessment. However, relatively little attention has been paid to it within the English qualifications context, perhaps because of the complexity of the concept of validity, or perhaps because, as some assessment researchers claim: 'it's all validity'. Abstracts in several other sections of this volume address validity issues, directly or indirectly. Abstracts in this section focus on the theory of validity and on how such a theory can and should inform the practical development and evaluation of qualifications.

Contemporary validity theory and the assessment context in England

Stringer, N. (2008)

The concept of validity underwent considerable development during the last half of the 20th century, from essentially meaning that 'a test measures what it says it measures', through multiple types of validity – content, criterion, and construct – to a multifaceted, but essentially unitary, concept of construct validity. The publishers of high-stakes tests in the US have, for the most part, embraced the modern concept of validity, while those responsible for general qualifications in England, such as the GCSE and GCE, appear not to have ventured far beyond evaluating content validity and reliability in ensuring the quality of these tests.

These differences may be attributable to differences between the two cultures, manifested in the form of tests and the personnel traditionally responsible for their construction. Nonetheless, the unitary concept of validity demands forms of evidence to counter threats to validity that content validity and reliability do not; as such, the quality of English general qualifications could benefit from explicit evaluation of validity, especially during specification (syllabus) development. The validity literature contains examples of the types of evidence required to satisfy validity concerns, and guidance on how to gather it. However, the involvement of the regulatory authorities in specification development means that responsibility for validity cannot lie exclusively with the awarding bodies, and a coordinated approach to validation would be required.

The achieved weightings of assessment objectives as a source of validity evidence

(Ofqual/14/5375)
Stringer, N. (2014)

In its proposed framework for validating Ofqual-regulated assessments, Ofqual identifies four key areas of evidence on which validity arguments should be based. One key area is the alignment between assessment and the curriculum/syllabus. The work reported here demonstrates a method for producing a source of validity evidence that falls under this category: the comparison of the intended and achieved weightings of assessment objectives at qualification and question-paper level. Screening data was produced at unit-subject level for the majority of AQA GCSE and A-level specifications, while six individual specifications were further analysed at the question-unit level. Where problems with achieved weightings occurred, some possible issues – both general to common assessment structures and paper-specific – were identified that could be fed back into the specification design and examination paper-writing processes. The complexities of interpreting and improving the achieved weightings of assessment objectives, in a context in which they are only one of a number of interrelated facets of validity, are discussed.

Fairness and differentiation

In England, the history of differentiation in examinations reflects changes in attitudes to both assessment and education. It was not until the Beloe Report in 1960 that assessment suitable for less academic students was considered. The report recommended a new system of exams (later to be known as the CSE) for those students who were not thought able enough to sit GCE O-levels. Even then, the new exams were only to be targeted at the second highest 20 per cent of students (with the assumption that O-levels targeted the top 20 per cent). In 1978, the Waddell report sowed the seeds for the more inclusive GCSE examinations; however, it was another 10 years before these exams were first taken. Research into differentiation in exams has to ensure that standards are equivalent across papers designed for different ability cohorts. It also has to address the (very topical) question of how a single exam can assess students from the whole ability range.

Assessments that differentiate successfully can contribute to fairness; however, they cannot guarantee it – there are other factors that can influence a student's attainment. Research that addresses fairness in assessment is often designed to reveal sources of unintended bias, as indicated by the abstracts below. Such research contributes to an understanding of the validity of assessments. Studies that consider the achievement of different socio-cultural groups, for example, might focus on 'face validity', which is concerned with the question of whether the context of an assessment item is equally accessible to all candidates.

Sex differences in GCE examination entry statistics and success rates

(*Educational Studies*, Vol. 6, No. 2)
Murphy, R. J. L. (1980)

Evidence is presented to show that entry patterns and success rates in GCE O-level and A-level examinations differ when the statistics for all subjects are combined, and, in different ways, when individual subjects are looked at separately. Furthermore, trends have been highlighted that reveal that various sex differences in examination statistics are gradually changing. A deeper understanding of these phenomena can only be arrived at by looking in more detail at particular aspects of the data: one example has been given to show how changes in the assessment techniques used in GCE examinations may directly contribute to sex differences in the results.

Some possible approaches to the problem of examining across a wide range of ability: A discussion of the new 16+ examinations

(*Curriculum*, Vol. 4, No. 2)
Cresswell, M. J. (1982)

This article discusses the strengths and weaknesses of a number of examination models that might be used to examine across a wide ability range. Some of the consequences of their use are outlined and it is clear that their appropriateness varies according to the nature of what is being assessed. It may be that where the emphasis of a syllabus is upon skills rather than mastery of specific content, it would be possible to set common papers consisting of questions that can be answered by candidates at a number of different levels. Hierarchical marking schemes could perhaps be employed to enable the anticipated wide range of responses to be scored.

Examination models involving alternative papers of differing difficulty might be useful where the range of complexity of the material to be covered in the examination is not too extensive, so that a central band of more than half the available grades is accessible to candidates taking either alternative. Where the provision of a reasonable overlap of grades between alternative papers is not possible – or the penalties attached to inappropriate entries in an alternative-papers model are judged to be too severe – a scheme involving one or two basic papers, and an optional extension paper leading to the higher grades, provides an attractive solution.

Can teachers enter candidates appropriately for examinations involving differentiated papers?

(*Educational Studies*, Vol. 14, No. 3)
Good, F. J. & Cresswell, M. J. (1988)

This paper considers the ability of teachers to enter candidates for appropriate combinations of differentiated papers. The results of experimental work suggest that teachers would be able to predict their pupils' examination performance accurately enough to enter almost all pupils at appropriate levels of such examinations; and that they would be able to do this as early as the January preceding the examination. However, they will only be able to enter candidates effectively if the standards required for the overlapping grades are the same at all levels of an examination. There is some evidence to suggest

that this condition may not always hold. In addition, results from some Joint 16+ examinations suggest that there may be a considerable number of inappropriate entries to GCSE examinations that use differentiated papers.

Setting common examination papers that differentiate

(*Educational Studies*, Vol. 15, No. 1)
Good, F. J. (1989)

Many GCSE syllabuses are assessed with examinations in which all candidates take the same papers. The setting of such papers is problematic because of the wide range of abilities and achievements of pupils at the age of 16, as well as the requirement that appropriate differentiation should be provided (i.e. opportunities must be given for candidates to show what they know, understand and can do). This paper considers a number of issues relevant to the setting of such examinations, including: how differentiation may be provided; the wording of questions; and how marks should be allocated. It highlights a number of potential pitfalls and concludes that although papers can be set that are accessible to all candidates and discriminate appropriately, common papers will not always provide adequate opportunities for the most able and least able candidates to show what they know, understand and can do.

What's in a name? Experiments with blind marking in A-level examinations

(*Educational Research*, Vol. 40, No. 2)
Baird, J. (1996)

A-level results have a substantial impact upon candidates' futures and it is crucial that the results are as fair as possible. Candidates' names appear on examination scripts and some have suggested that this could produce bias in the marking. Introduction of blind marking in A-level examinations would be unwieldy and costly. Two experiments on blind marking were carried out: one in A-level Chemistry, and one in A-level English Literature. In each subject, presentation (and not the content) of 30 scripts was varied. Eight Chemistry A-level examiners and 16 English Literature A-level examiners took part in the studies. Scripts were presented as blind or non-blind, with a male or female name and with 'male' or 'female' handwriting. The studies addressed the issue of possible gender bias in marking and investigated whether blind marking could overcome gender bias. It was concluded that

bias was not present in the marking and therefore no support was found for the introduction of blind marking in A-levels.

A feasibility study on anonymised marking in large-scale public examinations

Baird, J. and Bridle, N. (2000)

Names of centres were removed from summer 2000 GCSE and A-level examiners' stationery as part of a Department for Education and Employment initiative. Awarding bodies expressed concern about the possibility of mismatching candidates to their examination work if candidates' names were not visible on the examination script. A pilot study on anonymised marking was conducted in the summer 2000 AQA (SEG) GCSE English (2400PF) foundation tier examination. The purpose of this study was primarily to investigate whether administrative errors would occur due to the concealment of candidates' names. For two written-response question papers, candidates used examination booklets on which they wrote their name in the top right-hand corner and then concealed it from the examiner by folding and sticking the corner down. Over 34,000 candidates (68,000 scripts) were included in the study.

Teachers' views on tiering and ability grouping at GCSE

Baird, J. and Ireson, J. (2001)

Differentiation is handled in GCSEs by offering different tiers of entry for the examination. The typical assessment pattern is a higher tier, in which grades A* to D are obtainable, and a foundation tier, in which students can attain grades C to G. A questionnaire study with a sample of 50 teachers was carried out in each of the following GCSE subjects: English, French, History, Mathematics and Science (double). The questionnaires were followed by interviews with a small sample of teachers. Findings suggest that tiering does not drive ability grouping. However, the fact that many teachers allocate whole teaching groups to particular tiers (particularly in mathematics) suggests that ability grouping and tiering are associated. In contrast with findings in previous studies, teachers denied that the best teachers are allocated to any particular ability group. On average, few students changed ability group over the course of their GCSE study; and once they entered for a particular examination tier, few entries were changed. The combination of ability grouping and tiering could thwart the meritocratic intentions of our education system: with low expectations directly and indirectly serving to restrict opportunities for students who enter school with less social predisposition for educational success.

Assessment design

Our work on assessment design supports AQA in its development and delivery of qualifications, and also makes an important contribution to knowledge in a vital area. Research in assessment design spans the whole assessment process – from question-paper and mark-scheme design, through the marking stage, to the awarding of grades.

The format, and quite often the content, of assessments may have remained largely unchanged over the last 40 years, but the structure of examinations has evolved greatly; this has prompted intensive periods of research. Whenever a new mode of assessment is introduced, research is undertaken to investigate its validity. Ideally, the validity of any new design would be assured before it is released; however, such changes are often initiated by national policy decisions, leaving little time for validation.

The future of assessment design depends very much on the extent to which on-screen examinations take hold in England. While all assessments will soon be marked on-screen – and plenty of research has been undertaken to evaluate the effects of this (see pp. 46–53) – very few are currently taken on-screen. Whichever direction assessment takes, CERP's research activity will evaluate and inform its design.

Investigation into the relationship between grades and assessment objectives in History and English examinations

Joint Matriculation Board and the Schools Council
Orr, L. and Forrest, G. M. (1984)

This project aimed to obtain information on the mechanics of assessing pre-specified objectives, thus contributing to the process of devising a workable system of criteria-referenced grades. It followed government statements on the proposed single system of examining at 16+, which included two new grading scheme requirements: that there should be criteria referencing of grades, and that new grades should be linked with the present standards of CSE and GCE O-level grades. It was designed to highlight the way, and extent to which, attributes and skills stated in the syllabus as assessment objectives were reflected in the stages of the assessment process (question papers, mark scheme and scripts). The project included two subject-specific studies, History and English.

In both subjects, the examiners differed in their identification of assessment

objectives and their interpretations of explicitly stated assessment objectives. Both studies revealed that, often, not all the objectives that an examination question was designed to test were rewarded by the marks allocated. To conclude, it is recognised that the dependence of any assessment procedure on human judgement in assessment makes a measure of inconsistency unavoidable. To be compatible with a system of grade-related criteria, which are likely to be formulated in terms of fairly specific skills, there would be a need for the assessment objectives to be framed in much more specific terms than those that prevail at present. If criteria referencing of grades was to be adopted, examination questions would need to be more specific and all assessment objectives would need to be tested in each examination. In addition, mark schemes would need: to be clear and detailed; include hurdles in terms of specified attainments; restrict compensation between each of the relevant skills; and allow little, if any, room for idiosyncratic interpretation.

Describing examination performance: grade criteria in public examinations

(*Educational Studies*, Vol. 13, No. 3)
Cresswell, M. J. (1987)

At the time this paper was published, it was proposed to introduce grade criteria into GCSE examinations. One purpose of this move was to make more explicit the likely levels of competence and knowledge that might be expected from candidates obtaining particular grades. Grade criteria were also intended to provide realistic targets for teachers and pupils to aim at, and a consequential improvement in achievement was hoped for. It is argued that the former purpose is unlikely to be achieved at a more detailed level than that of a profile of a few sub-scores, or domains, within the subject. Further, achievement in each domain is likely to be reported in a way that permits only general interpretations: unequivocal inferences concerning candidates' specific achievements will not be possible. Finally, it is suggested that while the demands made of public examinations at 16+ are heterogenous and, in some instances, inconsistent, further progress on grade criteria is unlikely to be made.

Grading public examinations: the effect of contextual factors on grading outcomes

Houston, J. G. (1988)

This paper argues that contextual factors are crucial to any proper consideration of grading standards when these standards are determined

against subject-specific performance criteria such as grade descriptions. If contextual factors are ignored, comparability between the examinations of different authorities, say, or of one authority's examinations from one year to the next, is undermined. The paper highlights that it is difficult to quantify the effect of these contextual factors. All that those concerned with grading public examinations can do is to recognise their existence and make some allowance for them, providing there is agreement about whether a particular factor increases or decreases the demands of the examination.

Technical and educational implications of using public examinations for selection to higher education

(in Kellaghan, T. (1995) *Admission to Higher Education: Issues and Practice* [Dublin: Educational Research Centre])
Cresswell, M. J. (1995)

GCE A-level examinations are the principal selection device for higher education in England and Wales. This paper discusses some of the requirements that this use of the examinations places upon them, such as the apparent need for predictive validity, high-perceived reliability, and close comparability of standards between different A-level examinations. It is argued that these requirements have a considerable influence on the assessment techniques that are used, and that recent developments in England and Wales can be understood more easily if the role of public examinations in selection is taken into account.

Another difficult question? An investigation of problem solving and question-difficulty issues concerning gifted and talented students

Dhillon, D. and Richardson, M. (2003)

Two of the key design goals of the paper-based World Class Tests of problem solving are that they require the identification or construction of novel problem-solving strategies, and that they are cognitively demanding for the gifted students who take them. This paper investigates the problem-solving strategies that gifted students employ when answering such questions, and the effects on question difficulty of systematic manipulations to intrinsic cognitive demand and surface-level support or 'scaffolding'. Some insights into students' strategies were gleaned from qualitative interviews and script scrutiny, although these were hampered by students' poor meta-cognitive awareness. Manipulations to intrinsic cognitive demand had only a limited impact on students' performances. The effects of

scaffolding manipulations, although not significant, appeared to run counter to expectations; this suggests that provision of surface-level question support via structural, visual and strategy-cueing aids is a more complex task than anticipated.

Predictive models of question difficulty: A critical review of the literature

Dhillon, D. (2003)

Recent decades have seen a proliferation of research into the identification and manipulation of question-difficulty factors. This paper evaluates three predictive models of question difficulty, each of which provides valuable insights into some of the successes and potential pitfalls involved in the process of delivering an examination question at a specific level of difficulty. In the light of these insights, the viability of developing a unified model of question difficulty applicable across item types and subject domains is assessed. The paper concludes that for any such model to be useful, it may have to sacrifice a degree of rule-based precision in favour of flexibility and responsiveness.

Principles and practice of on-demand testing

(Report for Ofqual)

Wheadon, C., Whitehouse, C., Spalding, V., Tremain, K. and Charman, M. (2009)

This research was commissioned by Ofqual to review how advances in computer technology have been enabling on-demand testing in the UK, and to consider the implications of these advances for high-stakes general qualifications. The intention was to deepen understanding of the concerns of stakeholders in this area by looking at current practice in the UK and abroad. The first section of this report is a review of the literature relevant to on-demand testing. This review suggests that on-demand testing ranges from the provision of more frequent test windows to any time, anywhere testing. In its purest form, on-demand testing clearly supports the personalisation policy agenda and the desire to ensure all students achieve their potential.

In its less pure forms, the gains from on-demand testing include: increased efficiency in the assessment system, with more timely results; and flexibility in scheduling that frees the timetabling of the curriculum from fixed, arbitrary

examination dates. However, there are clearly risks inherent in redesigning the assessment system. Every process – from entries to results – is affected; these processes are complex and interlinked, and not all under the direct control of the awarding bodies. That said, no insurmountable technical difficulties were identified regarding issues such as the maintenance of standards over time, and between test versions.

The second section of the report details the views of some key stakeholders in the assessment process: teachers, students and examiners. The teachers and pupils were generally sceptical about the idea of pure on-demand testing supporting a personalised learning programme. They felt that this model would require support in terms of smaller class sizes and greater individual attention from teachers, for example, which would never materialise. Furthermore, both teachers and pupils were wary that an on-demand system would increase exam pressure through competition between peers and parents. However, they recognise that more flexibility in choosing testing dates could alleviate some existing pressures, as teachers would have greater control over the assessment timetable and therefore the delivery of the curriculum. The examiners welcomed the return to pre-testing that an on-demand system requires, and were generally positive about the assessment models that could be used to deliver on-demand testing in a rigorous manner.

The third section reports on a survey of current practice in on-demand testing. Nine major test providers supplied information regarding their current practice, either via interview or by responding to a detailed survey. The scale of provision is impressive. Hundreds of thousands of tests are being delivered on-screen, on-demand every year in the vocational and higher education arenas. Sophisticated technological infrastructures have been developed in partnership with technology companies. These partnerships are yielding innovative assessment formats based on realistic task-based assessments. However, there are few technology partners available, with eight out of the nine organisations surveyed sharing just two partners.

While it may seem that the major unitary awarding bodies are lagging behind in on-screen, on-demand testing, the concerns they need to satisfy are more complex. Vocational bodies tend to use a strong criterion-referencing standard setting approach that uses the judgement of experts to determine pass marks before tests are delivered. This can lead to large variations in pass rates over time, as seemingly superficial aspects of difficulty in a test can affect how candidates perform on them. This situation would not be tolerated in high-stakes national qualifications in England, not least because they are used as a benchmark for national performance. The awarding bodies would need more complex models of test delivery, with

statistical standard-setting models integrated into the test-construction and test-delivery processes, before they could countenance on-demand testing.

Finally, the report contains a draft set of principles for on-demand testing. These were initially drafted by the research team and presented to a group of technical experts who had substantial experience of working within UK awarding bodies and with on-demand systems operating outside the UK. Following their feedback, the principles were revised. While the experts broadly reached consensus on these principles, we do not claim that they are final and absolute. Rather, it is hoped that they will provoke discussion and debate, and lead to a rigorous framework within which on-demand testing can be regulated.

The value of AS-levels: increasing attainment and curriculum breadth or 'dumbing down' the gold standard?

Malpass, D. (2011)

This paper assesses the value of AS-levels by evaluating whether the qualification has achieved the aims that were set out when it was introduced as part of New Labour's Curriculum 2000 reforms. The first part of the paper considers whether the primary aims of AS-levels – to broaden the curriculum, and increase attainment rates by providing a more gradual gradient between GCSEs and A-levels – have been met. The second part of the paper explores the impact of AS-levels on teaching and learning, taking into account the views of students and teachers. It also considers the opinions of employers and Higher Education institutions that make decisions based on AS-level outcomes. The final section of the paper discusses the future of AS-levels, and considers whether the qualification should be retained in its current form, reformed to address its critics or replaced entirely with an alternative qualification. It concludes that AS-levels need to be reformed to enable candidates to develop effective study skills and an independent approach to learning. This will ensure that candidates are equipped with the necessary skills required for higher study, and that we provide a well-educated workforce to compete in the global marketplace.

Students and stakeholders

The policy context in which assessment takes place is continually changing, as is the way assessment is reported in the media, and hence understood by the public. CERP research seeks to identify the impact these factors have on assessment.

The position of teachers in relation to assessment has varied over time, as qualifications have ranged from comprising 100 per cent examination to 100 per cent coursework. These variations can have wide-reaching effects; research that turns the microscope on teachers' understanding of assessment is therefore vital.

The abstracts contained in this section illustrate that the views of all stakeholders need to be taken into consideration. Wider engagement can help the assessment community to fully understand the impact of its practices and, importantly, guide its efforts to increase public understanding of assessment.

Teachers' estimates of candidates' performances in public examinations

(*Assessment in Education*, Vol. 2, No. 1)
Delap, M. R. (1995)

Teachers' estimates of candidates' performance in public examinations are sometimes used as a trigger to investigate where the assessment procedures may have gone awry. In Britain, teachers' estimates are also used extensively in the selection of candidates to be interviewed and given conditional offers for entry into higher education. This paper presents multilevel analyses of over 7,000 estimated grades supplied in 1992. The results reveal that there are substantial differences between subjects. There is also some evidence to support the notion that teachers' estimates for females were slightly higher than for their male counterparts who obtained the same final grade.

Teachers' estimates of candidates' grades: curriculum 2000 A-level qualifications

(*British Educational Research Journal*, Vol. 31, No. 1)
Dhillon, D. (2005)

In the UK, estimated grades have long been provided to higher education

establishments as part of their entry procedures. Since 1994, they have also been routinely collected by awarding bodies to facilitate the grade-awarding process. Analyses of a British awarding body's data revealed that teachers' estimates of candidates' Curriculum 2000 A-level grades, in the first year of awarding, demonstrated an unprecedented degree of accuracy. Conversely, estimates of AS-level grades showed relative imprecision. Accuracy of the A-level estimates was most likely bolstered by feedback inherent to the modularisation of the examination, while the weakness of AS-level estimates may have been a consequence of the comparatively unfamiliar standard at which this qualification was set. As in previous research, when estimates were inaccurate they were more commonly optimistic than pessimistic (for both qualifications).

Media coverage of examination results, public perceptions, and the role of the education profession

Billington, L. (2006)

Examination issues have received increased media coverage over the past 30 years. This coverage has specifically focused upon examination results as a means of assessing examination standards. Each August, the debate regarding whether or not higher pass rates indicate falling or rising educational standards is played out in the media. This paper uses a sociological framework to examine the content of examination news items, public perceptions of examination standards, and the role of educators in the news coverage of examination issues. It is argued that an understanding of these issues is key to improving the annual news coverage of examination results. Equipped with such information, those responsible for educational assessment could develop effective strategies for interacting with the media.

A qualitative exploration of key stakeholders' perceptions and opinions of awarding-body marking procedures

Taylor, R. (2007)

The main aim of this study was to address a gap in the literature regarding perceptions of the examination system, by exploring perceptions and opinions of marking procedures among key stakeholder groups. Fourteen semi-structured interviews were conducted with teachers, parents, and examiners, and a focus group was conducted with five GCE students.

The findings suggest that parents, teachers and students are largely unaware of current marking procedures. In addition, there appears to be a lack of understanding of the examination system in general. Considering these findings, it is suggested that AQA should do more to increase the transparency of its routine processes, and should aim to increase understanding of the examination system among key stakeholder groups.

Stretch and challenge in A-level examinations: teachers' views of the new assessments

Baird, J., Daly, A. Tremain, K. and Meadows, M. (2009)

In recent years, some commentators have argued that A-levels do not stretch the most able students. To tackle such concerns, the government introduced a policy in the Education and Skills White Paper, which has become known as 'stretch and challenge'. New A-level examinations were introduced for first teaching in September 2009; these incorporated questions designed to induce more stretch and challenge in students' examination experiences. This research investigated teachers' perceptions of the new A-level question papers, and, more broadly, their experiences of preparing students for A-level examinations.

Engaging students via backwash: The A-level stretch and challenge policy

Baird, J., Chamberlain, S., Daly, A. and Meadows, M. (2009)

Internationally, governments have tried to increase students' higher-order thinking skills by changing examinations. Using the assessments in this way is an attempt at creating positive backwash effects. The case presented in this study is the English 'stretch and challenge' policy, which involves changes to the design of A-level examinations intended to produce higher-order thinking skills. This research investigated whether 39 students and 27 teachers considered that A-levels needed more stretch and challenge, and if they could perceive the intended policy changes in the new question papers. Findings were complex: while students and teachers agreed that the old-style A-levels were demanding, they considered them to be the wrong type of demands. Reported approaches to teaching and learning were highly strategic, with a great deal of emphasis placed upon studying past question papers and marking schemes. Students and teachers welcomed the stretch and challenge policy, and there were some indications that changes to the design of question papers could have the intended effects.

As this research was conducted prior to the introduction of the first new-

style stretch and challenge question papers, it is not possible to know whether positive backwash will occur. Research will need to be carried out when teachers have had the opportunity to adapt their teaching practices to the new demands.

Communication strategies for enhancing qualification users' understanding of educational assessment: recommendations from other public interest fields

(*Oxford Review of Education*, Vol. 39, No. 1, *Special Issue: The public understanding of assessment*)
Chamberlain, S. (2013)

In many countries, the outcomes of national assessments provide 'qualifications' or 'credentials' that may be used to define the levels of students' knowledge and skills. These definitions are relevant for employers, higher education institutions and others. Qualification users – such as students, parents and teachers – arguably need to have an understanding of some basic principles of educational assessment in order to make informed judgements about the reliability of assessment outcomes; they also need to develop realistic expectations of what assessment systems can deliver. Endeavours to achieve this have gathered pace recently, with the completion of a two-year research programme in England that explored concerns around technical aspects of assessment and current levels of public understanding of assessment.

One of the recommendations of the programme was that awarding bodies should collect, and make available, information relating to the reliability of outcomes for various types of qualification. However, further consideration is required to determine what, and how much, assessment information would be useful to qualification users; and how it might best be presented and disseminated. This paper discusses the communication strategies employed in other fields for the purpose of sharing important messages with the public. Three recommendations are offered for overcoming some of the challenges inherent in improving the communication and understanding of assessment.

The paper concludes that enhancing qualification users' understanding of assessment may be achieved by focusing on the presentation of applied, interpretive information, and disseminating it through influential peers from various stakeholder groups.

Ripping off the cloak of secrecy

Professor Paul Newton gave a keynote speech at an event held as part of the AQA Research Committee anniversary commemorations. Alex Scharaschkin, Director of the Centre for Education Research and Practice (CERP), was the respondent. Professor Jannette Elwood chaired the event, which was held at Kings Place, London, on 1 December 2015. A transcript is reproduced below

Paul Newton, Research Chair, Ofqual

Without wishing to become unduly existential tonight, the reason that I'm here today – as an assessment researcher – can be traced back to the summer of 1994, when the Associated Examining Board (AEB) offered me my very first job. So I'm pleased to have been invited to say a few words at this forty-year celebration of the AEB (now AQA) research group, on the year of my own coming of age.

I'm going to talk about the role of research in ripping off the cloak of secrecy, with particular reference to research conducted within AQA and predecessor bodies – AEB, Northern Examinations and Assessment Board (NEAB), and the Joint Matriculation Board (JMB). Of course, other brands are available – and I'll say a few words about them too. But I'll start with a letter to the editor of *The Lancet*:

'SIR, – Your attention has been directed to abuses existing in the government of the Veterinary College, and especially to the mode in which the examinations are conducted.

That government already trembles, and its members, like cattle affrighted at the distant rolling of the thunder storm, *have assembled together* to seek the means of evading the dreaded effects, to avoid, by enveloping their proceedings in the cloak of secrecy, the searching shaft of scrutiny, the flash of discovery, and the drenching shower of deserved obloquy.'
(Beard, 1826)

That was published in 1826. While the language may sound a little dated, the 'cloak of secrecy' critique still resonates, even today. As exam boards – and regulators – we often feel like the target of public criticism. Sometimes it feels like our stakeholders don't entirely trust us. Unfortunately, to the extent that that is true, it's a real problem; not just for us, but for society. That's because examination systems – in comparison with many other systems – are supposed to be Whitest of All. And that brings us to the 'Daz Challenge'.

To a large extent, our public examination systems are a response to the refusal of society to accept historical traditions of nepotism, patronage and corruption when it comes to allocating and withholding valuable educational or social opportunities – like places at university or jobs in the civil service. For that reason – an underlying distrust of people in power – we don't just want our examination systems to be fair, we want them to be translucently fair. We actually want to *see* that fairness. We want to do the 'Daz window test' on our public examinations and see the fairness shine through. And why wouldn't we? Examinations play a huge role in structuring the society that we live in; we have to be able to trust them.

Strangely, though, our examination systems seem to have been 'surrounded with mystery' (Wallis, 1927, p. 2) even from the outset. Bertie Wallis – who was a chief examiner of sorts – published a book in 1927 that identified six problematic traditions in examining, including secrecy. He also provided a really interesting diagnosis of why examination systems may seem to be so secretive. Firstly, examiners hate to be unfair; however, they recognise the impossibility of complete fairness, owing to the inevitability of imprecision and error. Wallis suggests that this tempts them to conceal their activities. Secondly, if you don't fully understand how an examination system works, then it's much harder to complain about it – so secrecy can be used to disempower customers and members of the public. Thirdly, and more malevolently, secrecy can be used to conceal bad examining practices. Finally, secrecy can provide an excuse for not improving poor examining processes. Basically, Wallis thought that the tradition of secrecy was a bad thing and needed to be broken down.

Within a year of the publication of Wallis's book, James Murray Crofts produced *Secondary School Examination Statistics*. Crofts was the secretary to the JMB from 1919 to 1941. And he was clearly committed to improving public understanding of the examination system. This is what he said in the introduction to his book:

'We have moved some little way from the haphazard methods of conducting examinations which were customary in bygone days, but it is clear that … the working of the whole machine is [still] imagined to be soulless, conscienceless, unthinking, and unintelligent.

Though the conduct of examinations is a matter for the expert, it is right, too, that even the general public should know in outline how they are conducted.' (Crofts & Caradog Jones, 1928, pp. v–vi)

So he was a real pioneer of promoting public understanding. From the 1920s onwards, the JMB championed a new approach to examining.

This approach was (according to Crofts) neither soulless nor conscienceless, because it was driven by the professional expertise of subject examiners, and neither unthinking nor unintelligent, because it was moderated by the new science of statistical reasoning.

And, thus, the cloak of secrecy was ripped off? Er, no! This is how a subsequent JMB secretary – James Petch – put it, 35 years later:

'There has previously not been any reluctance on the part of the Board to answer fair enquiries reasonably made by responsible folk.

But on the whole it has preferred to get on with the job and not to proffer comment or information where it has not been asked for.'
(Petch, 1963, p. 3)

In other words, somehow – between the pioneering work of the second secretary (James Crofts) and the equally important work of the fourth secretary (James Petch) – neither the JMB nor any of the other boards seemed to do anything much at all to break down this longstanding tradition of secrecy; with one or two occasional exceptions. In fact, it wasn't really until after the Second World War that the boards began to get serious about ripping off the cloak of secrecy, through books written by the following authors:

Brereton (1944)
Petch (1953)
Jeffery (1958)
Wiseman (1961)
Bruce (1969)

And what about the role of research in ripping off the cloak of secrecy? Well, this is what Stephen Wiseman said in 1961, in relation to reliability research:

'What is disturbing, however, is the paucity of any published inquiries of this kind sponsored by the examining boards themselves.'
(Wiseman, 1961, pp. 139–40)

And this is what he said in relation to comparability research:

'[If the boards had published their own research, it] would have helped to prevent the spread of horror-stories about such things as lack of equivalence which is an inevitable concomitant of the present cloak of secrecy.'
(Wiseman, 1961, p. 154)

So this leads me to another interesting question. Why does there seem to be so little evidence of exam board research activity from the 1930s to the 1960s, particularly as most of the boards were located within universities? I have two hunches.

My first hunch is that there may have been an element of diffused responsibility for examinations research. A primary role of the Secondary School Examinations Council – which ran from 1917 to 1964 – was to inquire into comparability. It didn't publish many investigations, but it did publish some – including two major reports in the 1930s. So maybe this diffused responsibility let the boards off the hook to some extent?

My second hunch is that any early appetite for conducting and publishing research into examinations would have been completely lost in the wake of the International Institute Examinations Inquiry. This inquiry was funded largely by the Carnegie Corporation, in the US – at least partly to proselytise North American research and development in the field of objective testing. Some of the leading scientists of the day were appointed to the English Committee – including Cyril Burt, Godfrey Thomson, Charles Spearman, and so on – some really big hitters! Even before the inquiry had begun, these committee members held some really strong views on examinations. This is how Philip Hartog understood the mission of the inquiry:

'I am perfectly certain that in England you can only enter into the citadel of examinations as they are now, blow up what is bad and reconstruct what is good, you can only enter a citadel with a battering ram of facts, and it is as such a battering ram that I regard the preliminary enquiry which we have sketched out.'
(Hartog, as cited in Lawn, 2008, p. 39)

The inquiry basically involved loads of marking reliability studies – across all sorts of examinations – and came up with conclusions like this (for School Certificate History):

'No process of measurement can be valid when it yields such discrepant results in the hands of the same two examiners on two different occasions.'
(Hartog & Rhodes, 1935, p. 16)

The inquiry got massive coverage in 1935 when *An Examination of Examinations* was published – making the exams crises of 2002 and 2012 look a bit like a walk in the park. And, although the boards (and others) strongly protested that the methods were flawed, the credibility of examinations during this period was severely damaged. In other words,

instead of ripping off the cloak of secrecy, this publication probably caused the cloak to be wrapped even tighter.

It took a long time before the exam boards began to talk openly about the research that they were conducting; so it's really hard to know whether there was much going on at all. But this is a quotation from a chapter by J. G. Jenkins, who was secretary of the London Board:

'The question of comparability of standards is … so important … that the examining bodies themselves have seen fit to institute joint investigations of their respective standards in selected studies.'
(Jenkins, 1958, p. 128)

So, the boards must have been conducting some research during the 1950s, but they were still extremely reticent to publish it. In fact, it wasn't until the 1970s that they began to publish in earnest. This is how they explained their reluctance to publish, in the first review of GCE comparability studies:

'In presenting this booklet to the public … we in the GCE boards have found ourselves in a dilemma. If we merely state that comparability exercises are regularly conducted and do not show our hand, we appear to have something to hide. If we try to explain them, their complexities and limitations invite misunderstanding and misrepresentation. On balance, the preferable alternative seemed to be to "publish and be damned". We have, and probably shall be.'
(Robin Davis, as cited in Bardell, Forrest & Shoesmith, 1978)

Many of you will be familiar with this report, and you'll recognise that particular passage. But did you know about this report: *A Review of Investigations into Subjects at Advanced Level Conducted by the G.C.E. Boards: 1953-1968*? This is the *real* first review! As far as I'm aware, it wasn't published. My copy has a note, in handwriting, on the front cover that says: 'January 1970 revision, for forwarding to the Schools Council.' Again, on my copy, the final paragraph says this:

'It is regrettable that publicity has not been given to this work, but many of the findings were only of concern to the boards involved, and quotations and statistics taken out of context are notoriously liable to misinterpretation, or misconstruction.'

More interestingly, perhaps, on my copy, this final paragraph has been crossed out. Clearly, even in 1970, the boards felt unable to publish this kind of work – as useful to them as it might have been to publish it. But that was

soon to change – and exam board research was soon to play a major role in ripping off the cloak of secrecy.

Now, although it is possible to find evidence of research activity during the 1930s, and even during the 1800s, it was during the 1960s that the boards began to take research seriously. The JMB established a Research Unit in 1964. The Schools Council established the Examinations and Tests Research Unit within the NFER in 1965 (which ran until 1977). A cross-board Test Development and Research Unit was established between UCLES, OCSEB, UODLE in 1967 (which ran until 1985). The GCE Standing Research Advisory Committee was established in 1970 (which represented all of the boards, including AEB). The AEB Research Group was formed in 1975. And the rest, as we like to say, is history. It's a very strong history, which led to many powerful publications...

...from the JMB, for example:
Forrest and Smith (1972) – one of the first statistical analyses of inter-subject comparability.
Whittaker and Forrest (1983) – a highly influential paper on the problems of interpolating by percentages.

...from the NEAB, for example:
Jones (1997) – on the ambiguity of statistical analyses of comparability.
Fowles (1995) – one of the many inter-board comparability studies.

...from the AEB, for example:
Murphy (1978) – 40 years after Hartog and Rhodes, one of the very first studies of marking reliability by an exam board researcher.
Good and Cresswell (1988) – they've got an effect named after them; enough said!

...from AQA, for example:
Baird and Scharaschkin (2002)
Meadows (2012)
These final two examples provide a name check for the three most recent research directors: Jo-Anne (Baird), Michelle (Meadows) and Alex (Scharaschkin).

Of course, there were many publications from the other boards and organisations, too. Incidentally, the fact that Cambridge Assessment is represented twice in this presentation isn't expressing any regulatory preference. The second publication is by my wife – so that's just pure and simple nepotism.

So, the cloak of secrecy… has it been removed? Well, I think the exam boards would *like* to think so. And I think the regulator would *like* to think so. But what about everyone else? Just a few months ago, the HMC (Headmasters' and Headmistresses' Conference) chair, Richard Harman, said:

'One way to reduce this problem for future years would be through less secrecy and greater transparency; via publication of re-grade statistics by subject and by exam board. But the real remedy lies in more accurate first-time marking.'
(Harman, 2015)

I'm not presenting this quotation either to support it, or to challenge it, or to open it up for debate here. That's a discussion for different people on a different occasion. I'm just inviting you to reflect on its theme – marking reliability – exactly 80 years after the publication of *An Examination of Examinations*, by Hartog and Rhodes. And I'm just observing that the secrecy narrative can still be heard. We've undoubtedly come a long way in terms of openness – and exam board research has taken us forward in leaps and bounds – but we haven't entirely arrived yet. In other words, there's still a lot of work for AQA researchers to do! So this is where I'll end, with a succinct conclusion.

Our examination system is, I genuinely believe, fairer than it's ever been. And, to a large extent, we have decades of exam board research to thank for that. The system is also more open. And decades of exam board research publications pay testament to that fact.

However, our examination system is still not entirely translucent. To be fair, it could probably never be as translucent as some of our stakeholders would like it to be. Because it's a deceptively complex system: it looks like it ought to be simple to demonstrate its fairness; but it really isn't anything like simple to do so. And that's our challenge – the Daz challenge – which requires a 'Bold' response from exam boards and regulators alike.

So, to round it off, there's nothing else I need to say other than: Keep Calm and Carry on Researching!

References

Baird, J., & Scharaschkin, A. (2002). Is the whole worth more than the sum of the parts? Studies of examiners' grading of individual papers and candidates' whole A-level examination performances. *Educational Studies*, *28*(2), 143–162.

Bardell, G. S., Forrest, G. M., & Shoesmith, D. J. (1978). *Comparability in GCE*. Manchester: Joint Matriculation Board.

Beard, J. (1826, July 15). Abuses existing in the government of the Veterinary College [Letter to the editor]. *The Lancet*.

Brereton, J. L. (1944). *The Case for Examinations: An account of their place in education with some proposals for their reform*. London: Cambridge University Press.

Bruce, G. (1969). *Secondary School Examinations: Facts and commentary*. Oxford: Pergamon Press.

Crofts, J. M., & Caradog Jones, D. (1928). *Secondary School Examination Statistics*. London: Longmans, Green and Co.

Forrest, G. M., & Smith, G. A. (1972). *Standards in Subjects at the Ordinary Level of the GCE, June 1971*. Manchester: Joint Matriculation Board.

Fowles, D. (1995). *A Comparability Study in Advanced Level Physics: A study based on the summer 1993 and 1990 examinations*. Manchester: Northern Examinations and Assessment Board on behalf of the Standing Research Advisory Committee of the GCE boards.

Good, F. J., & Cresswell, M. J. (1988). Placing candidates who take differentiated papers on a common scale. *Educational Research*, *30*(3), 177–189.

Harman, R. (2015, August 10). 'Name and shame' exam boards who fail to mark A-level papers properly. The Telegraph. Retrieved from http://www.telegraph.co.uk/education/educationopinion/11792861/Name-and-shame-exam-boards-who-fail-to-mark-A-level-papers-properly.html

Hartog, P., & Rhodes, E. C. (1935). *An Examination of Examinations*. London: Macmillan and Co.

Jeffery, G. B. (Ed.). (1958). *External Examinations in Secondary Schools: Their place and function*. London: George G. Harrap & Co. Ltd.

Jenkins, J. G. (1958). Operation GCE. How the examination is conducted. In G. B. Jeffery (Ed.), *External Examinations in Secondary Schools: Their place and function* (pp. 112–128). London: George G. Harrap & Co. Ltd.

Jones, B. E. (1997). Comparing examination standards: is a purely statistical approach adequate? *Assessment in Education: Principles, Policy & Practice*, *4*(2), 249–263.

Lawn, M. (2008). Blowing up the citadel of examinations: the English Committee and the Carnegie Corporation. In M. Lawn (Ed.), *An Atlantic Crossing? The work of the International Examination Inquiry, its researchers, methods and influence* (pp. 39–59). Oxford: Symposium Books.

Meadows, M. (2012). *From Benchmark to Judge's Bench: An insider's view of the causes of the 2012 GCSE English exams crisis*. Manchester: Centre for Education Research and Policy.

Murphy, R. J. L. (1978). Reliability of marking in eight GCE examinations. *British Journal of Educational Psychology*, *48*(2), 196–200.

Petch, J. A. (1953). *Fifty Years of Examining*. London: Harrap & Co.

Petch, J. A. (1963). *The Joint Matriculation Board: What it is and what it does. Occasional Publication 16*. Manchester: Joint Matriculation Board.

Wallis, B. C. (1927). *The Technique of Examining Children: A Quest for Capacity*. London: Macmillan and Co.

Whittaker, R. J., & Forrest, G. M. (1983). *Problems of the GCE Advanced Level Grading Scheme*. Manchester: Joint Matriculation Board.

Wiseman, S. (1961). The efficiency of examinations. In S. Wiseman (Ed.), *Examinations and English Education* (pp. 133–164). Manchester: Manchester University Press.

Additional sources

Cooke, G. (2008). Research and development. In S. Raban (Ed.), *Examining the World: A history of the University of Cambridge Local Examining Syndicate* (pp. 158–178). Cambridge: Cambridge University Press.

Cresswell, M. J. (2000). *Research Studies in Public Examining*. Guildford: Associated Examining Board.

James, H. (2003). The Joint Matriculation Board and the Northern Examinations and Assessment Board. In H. James (Ed.), *Setting The Standard: A century of public examining by AQA and its parent boards* (pp. 55–84). Manchester: Assessment and Qualifications Alliance.

Kingdon, M. (2007). Commentary on Chapter 2. In P. E. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 92–94). London: Qualifications and Curriculum Authority.

Tattersall, K. (2007). A brief history of policies, practices and issues relating to comparability. In P. E. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 43–91). London: Qualifications and Curriculum Authority.

Willmott, A. S. (1980). Twelve Years of Examinations Research: ETRU 1965-1977. London: Schools Council.

Alex Scharaschkin, Director, CERP

The Daz challenge provoked a couple of thoughts I'd like to share (perhaps by way of a kind of pre-wash cycle to the wider discussion we will have in a moment).

It was interesting to consider the influence of Hartog and Rhodes' work in the first half of the 20th century. Their aim was, largely, to replace forms of assessment traditionally valued in the English system, such as essays, with objective tests. (This was certainly an aim of Cyril Burt, who wanted to replace curriculum-embedded examinations with IQ tests). Their argument, as Paul said, was essentially couched in terms of reliability – and specifically, in terms of marking reliability, or what we might call these days the 'rater effect' component of overall reliability. The solution to the existence of such effects was to abolish any questions that would require the exercise of examiner judgement in their marking. Essays out: multiple-choice questions in.

Now, objective-test-style and multiple-choice questions did, of course, become part of the repertoire of assessment options in English public examinations during the 20th century, and continue to retain an important role. And yet 'traditional' assessment tasks, such as questions requiring discussion, or more open-ended problems in which candidates have to develop their own approaches and arguments, did not wither away, as was perhaps hoped for by the proselytisers of US-style objective testing in the '30s. They stubbornly kept their places in the examiners' armoury – and have done, to this day.

Indeed, there will be more extended prose, for example, in the new GCSE and A-level qualifications that will be examined from 2017 onwards. I think this means that, as a society, we believe that valid assessment of the kinds of skills pupils need to prepare them for advanced study, or to deal with the world of work, must involve an element of 'showing what you can do': in other words, constructed responses such as extended writing, performance, drawing, or speaking.

This brings me on to research – and AQA's research in particular – and its role in addressing the Daz challenge.

If we value this kind of assessment – that is, if we wish to retain assessment styles that require a human marker or grader – then we need to further explore examiner judgement. What are its advantages; what are its potential limitations? How do we embed the former and mitigate the latter in our approach to assessment design? How do we build consistency of understanding of what 'good performance' – or performance at a particular level or grade – looks like in the various domains our assessments cover? By investigating these issues, and acting on the results, we have the best chance of ensuring that our assessment procedures are as fair as they possibly can be: that they maximise translucency, in Paul's terms.

That is why the issue of examiner judgement, in its broadest sense, has been a key theme in the research carried out by AQA over the years. Our research has covered topics such as the cognitive demands involved in recognising whether, for example, two examination performances are of the 'same standard', when the questions that prompted them were different (a task involved in maintaining standards in examinations between years). It has examined the effects of context on perceptions of quality of performance on a question. We have looked at the extent to which experts are able to predict the demand level of particular questions. We have studied how experts measure performances against a set of marking criteria, and how well the criteria enable them to categorise performances consistently by value.

If you look at the mark schemes that are used these days to differentiate between, and appropriately reward, students' responses to GCSE and A-level questions, and compare them with those of a decade or two ago, you will see the fruits of this research. You will also see it in the improvements in statistical measures of marking reliability, and other quality of marking metrics, over time.

Much of this research has been done by people who are in the room this evening. Paul has referenced a few of the AQA papers – published in peer-reviewed journals, and available in the public domain – that have had a significant influence on assessment design, delivery, and policy in this country over the last 40 years. There are many others one could point to, if time permitted. And of course the work continues. Currently we are researching, among other topics:

- the application of comparative judgement in assessment, and its cognitive basis
- the reliable classification of multifaceted, constructed responses
- the nature of examination standards themselves, and how our conceptions and constructions of 'standards' compare with those in other countries
- how to temper known biases or limitations that arise from the unsupported use of judgement, with statistical evidence, to ensure comparability of grading standards between years
- how to design marking schemes for maximum effectiveness, and how to use e-marking technology to best effect to monitor quality of marking in real time.

So, I wholeheartedly endorse Paul's conclusion that research carried out by AQA – and the other exam boards – has greatly improved the situation since Hartog and Rhodes' day. Not only do we know far more about how to design and administer assessments that properly balance validity and reliability, we have, I believe, well and truly ripped off the cloak of secrecy.

Ultimately, perhaps, the aim of assessment regimes is to tell us about things that we, as a society, regard as valuable outcomes of education for our young people. As our *values* – and hence what is *valid*, in assessment – evolve and change over time, so our assessment regimes must evolve alongside them.

In part, that may involve technological change. Will future assessment all be done online, using simulations and complex, adaptive learning environments? Quite probably. As to the timescale for that, I think it is probably one of those socio-economic phenomena that Rudi Dornbusch characterised when he said that 'in economics, things take longer to happen than you think they will, and then they happen faster than you thought they could.'

More fundamentally, what we as a society think is important and valuable in education will no doubt also evolve. There is clearly a balance to be struck between assessing knowledge, and assessing skills. It is a question of curriculum design (rather than strictly of assessment) as to what/how much 'core knowledge' is right. Over and above that, though, when we at AQA talk to employers, and to higher education providers, it is clear that as well as knowledge, they value skills such as problem solving, creativity, adaptability, teamwork and perseverance.

Such things are hard to assess – but we have to keep trying. And I think our experience in dealing with some of the thorny questions that arise in the 'English' style of examination assessment actually stand us in good stead for this. One piece of evidence for this is the requests we have had from the Chinese government to introduce our A-level-style assessment in high-performing schools in China, because they believe it prepares students more effectively in just these kinds of areas.

So, it is really important that, at AQA, we continue research in all the areas I have touched on (and many that, for reasons of time, I haven't). By doing so, we contribute to and support the enduring charitable purpose of AQA, which is to help students and teachers realise their potential.

I'd like to thank all the researchers – past and present – who have contributed to that aim, and whose research (in many cases pioneering and far-sighted, and always conducted to the highest standards of academic rigour) we are celebrating this evening. I'd like to thank Paul, once again, for such an excellent lecture. I'm not sure that I can promise that we'll always be able to keep calm, but I can assure you that we will carry on researching.

Formation of the Centre for Education Research and Practice (CERP) and the AQA Research Committee: A timeline of key events

1903 The Joint Matriculation Board of the Universities of Manchester, Liverpool, Leeds, Sheffield and Birmingham (JMB) is formed.

1953 The Associated Examining Board (AEB) is formed.

1964 The JMB establishes its Research Unit (RU). J. A. Petch is appointed director as its director.

1965 The RU establishes a committee to oversee its research. The Certificate of Secondary Education (CSE) is introduced.

1968 Gerry Forrest is appointed director of the RU.

1973 The RU forms the Research Advisory Committee (RAC).

1975 The AEB establishes its Research and Statistics Group at Aldershot, led by Jim Houston.

1985 AEB's research unit relocates from Aldershot to Stag Hill House, Guildford.

1988 GCSEs first examined.

1990 A. M. Spencer appointed director of the RU.

1991 Mike Cresswell appointed Head of Research at AEB.

1992 JMB and the Northern Examining Association merge to form the Northern Examinations and Assessment Board (NEAB). The RAC is dissolved.

1994 The AEB assumes control of the Southern Examining Board (SEB), although both boards retain their respective identities.

1997 AEB/SEB, NEAB and City and Guilds begin collaborative work under the newly formed Assessment and Qualifications Alliance (AQA).

2000 AEB/SEB and NEAB merge to form a single awarding organisation known as the AQA. The AQA Research Committee is created.

2001 Curriculum 2000 modular AS qualifications first certificated.

2002 Curriculum 2000 modular A-level qualifications first certificated. Jo-Anne Baird is appointed as AQA's Director of Research.

2004 Cresswell is appointed CEO of AQA.

2007 Stephen Sharp is appointed Director of Research.

2008 Michelle Meadows is appointed Director of Research.

2009 Jannette Elwood is appointed as chair of the AQA Research Committee.

2010 Andrew Hall is appointed CEO of AQA.

2011 AQA's research department is rebranded as the Centre for Education Research and Policy (CERP).

2014 CERP is renamed the Centre for Education Research and Practice (CERP).

2014 Alex Scharaschkin is appointed Director of CERP.

2015 Scharaschkin is appointed Director of Research and Compliance.

The Research Committee

The AQA Research Committee is a prestigious advisory group of national and international researchers. All research carried out by AQA's Centre for Education Research and Practice (CERP) undergoes an exacting peer review process, and the committee meets twice a year. The current committee can trace its roots back to the Associated Examining Board (AEB) and Joint Matriculation Board (JMB) advisory groups (see pp. 8-17). The following individuals have contributed to the various incarnations of the committee over the years and ensured that the research outlined in this publication has the appropriate credibility and academic rigour:

Mr Adams
Dr A. Ahmed*
Professor J. T. Allanson
Mr W. F. Archenbold
Mr Archer
Mr Adams
Professor J. Baird*
Mr D. Battye
Dr A. Beguin
Dr Black
Professor Blackman
Mr Booth
Dr R. J. Bradbury
Dr Buckland
Professor Burkhardt
Mr G. Carver
Mrs D. Chambers
Mr P. Charles
Mr T. Christie
Professor D. Clark-Carter
Professor R. Coe*
Mr Crow
Dr B. Crowther
Mrs C. De Luca
Dr Dobson
Professor Driver
Mr Duffin
Professor Eggleston
Professor J. Elwood*
Dr A. Feiler
Mr Fitzgerald
Dr Fitz-Gibbon
Dr French
Professor S. French
Mr G. Glyn
Professor Gosden
Dr C. Gray
Mr Gray
Mr Greenwood
Miss Hardcastle
Professor J. Harding
Mr Hathaway
Professor Heathcote
Mr P. Hendry
Professor Holliday
Professor P. Huddlestone*
Dr W. D. Ions
Dr T. Isaacs*
Mr J. Johnson*
Mr Kirshner
Dr S. Knutton
Mr D. Linnell
Mrs Livesey
Mr Locke
Mr Mathews
Mr Mathieson
Miss Moore
Mrs S. Moore*
Mr G. E. Mountfield
Mr T. Mullen
Professor R. J. L. Murphy
Dr T. Myers
Dr C. A. Newbould
Professor P. Newton*
Dr D. Nicholls
Mr Ogborn
Mr Ogden
Professor Oliver
Mr A. Pollitt*
Mrs N. Powrie*
Professor P. F. W. Preece
Mr Rayner
Mr Reid
Mr L. Ridings
Mr Robinson
Mrs S. Rogers
Mr Rowlands
Mr Sanders
Miss Sands
Dr I. Schagen
Mr A. Scharaschkin
Mr Sharp
Dr Sheppard
Mrs A. Smith
Mr Starr
Dr G. Stobart
Mrs J. Sturgis
Ms Styles
Professor S. Strand
Ms Sutton
Ms Tattersall
Dr R. Taylor
Mr Turner
Professor P. Tymms
Mr G. van Lent
Dr Verma
Mr Viner
Mr N. Walkey
Miss Whittaker
Mr E. R. Whitworth
Dr C. Wikstrom
Professor D. Wiliam
Professor A. Wolf
Ms S. Wright*
Mr G. Young

*Indicates current AQA Research Committee member

The Centre for Education Research and Practice

Research and technical staff as of 2015:

Alex Scharaschkin (Director of CERP)
Anton Béguin (Director of Research and Innovation)
Lena Gray (Head of Research)

Emma Armitage
Yaw Bimpeh
Simon Eason
Liz Harrison
Ruth Johnson
Ben Jones
Kate Kelly
Lesley Meyer
Caroline Paget
William Pointer
Ben Smith
Victoria Spalding
Charlotte Stephenson
Daryl Stevens
Neil Stringer
Zeek Sweiry
Martin Taylor
Claire Whitehouse
Alison Wood

Centre for Education Research and Practice (CERP)

AQA's Centre for Education Research and Practice (CERP) provides robust evidence that informs both organisational direction and wider educational debate.

CERP is a multi-disciplinary research facility with sites in Manchester, Guildford and London. We have a record of high-calibre research that stretches back 40 years through our predecessor bodies, as indicated throughout these pages. While we maintain exacting standards of academic rigour – our work is reviewed by a prestigious committee of national and international researchers, chaired by Jannette Elwood of Queen's University, Belfast – current research is grounded in the practical realities of assessment and qualifications.

Our team comprises statisticians, psychologists, educationalists and scientists. These varied backgrounds allow us to employ a range of qualitative and quantitative methodologies, which ensures our findings are credible and valid.

We are passionate about education, particularly high-stakes qualifications, and we share our work with a wide audience that includes the academic community, policy-makers, teachers and specialist media. We regularly publish research papers, blogs and longer articles via our website (cerp.org.uk), which can be freely accessed.

Examining assessment

AQA's research – carried out within the Centre for Education Research and Practice (CERP), under the auspices of the AQA Research Committee – has its roots in a number of predecessor bodies, including the Joint Matriculation Board (JMB) and the Associated Examining Board (AEB). This publication highlights the research that has been undertaken during the last four decades, and marks the 40th anniversary of the research committee.

CERP, together with its predecessor research units, has produced hundreds of reports. This compendium summarises that significant body of work, and reflects on the changing face of assessment research.

Most of the papers cited here can be read in full on our website at **cerp.org.uk**.

For more information, contact **cerp@aqa.org.uk**.

AQA
Realising potential