What happened to limen referencing?
An exploration of how the Awarding of public examinations has been and might be conceptualised

Jackie Greatorex

A paper presented at the British Educational Research Association Conference, 10 to 13 September 2003 at Heriot-Watt University, Edinburgh

Disclaimer
The opinions expressed in this paper are those of the author and are not to be taken as the opinions of the University of Cambridge Local Examinations Syndicate (UCLES) or any of its subsidiaries.

Contact details
Jackie Greatorex, RED, UCLES, 1 Hills Road, Cambridge, CB1 2EU.
greatorex.j@ucles.org.uk.

Available at www.ucles-red.cam.ac.uk

What happened to limen referencing?
An exploration of how the Awarding of public examinations has been and might be conceptualised

1 Abstract

It is generally accepted that public examinations in England and Wales are neither criterion referenced nor norm referenced. The Awarding process where grade boundaries are chosen is characterised by combining professional judgements of quality with statistical evidence to determine grade boundaries. A grade boundary is the minimum mark that a candidate must achieve to be awarded a given grade. In the Awarding process the judgementally determined grade boundaries are the grade A and E boundaries at A level and the grade C, A and F boundaries for nontiered GCSEs. The remaining grade boundaries are determined arithmetically. The Awarding process has been conceptualised by different authors as limen referencing, cohort referencing and weak criterion referencing.

There is little literature about limen referencing. When the Awarding process has been conceptualised as limen referencing an A grade boundary is argued to be the limen (threshold) at which the unique qualities of A grade achievement are detectable, and they are undetectable below the grade boundary. The same principle is applied to the other judgementally awarded grades. Cohort referencing is ensuring that the proportion of candidates awarded each grade remains the same from one examination session to the next. Weak criterion referencing is relaxing the notion of criterion referencing, it involves *maintaining the general quality of examination performance required for each grade, given the difficulty of the examination but not demanding evidence of specific achievement* (Baird *et al*. 2000, 215).

In this paper the Awarding process and associated concepts are revisited with the aim of scrutinising the different terminology that has been used to describe Awarding and to consider how it might be conceptualised.

2 The Awarding process

Prior to the Awarding Meeting for a General Certificate of Education (GCE)[i] or General Certificate of Secondary Education (GCSE)[ii] the Principal Examiner for an examination paper will have written the paper, marked some (or all) of the scripts and supervised all additional marking. S/he will use this experience and his/her experience of grade boundaries in previous years to suggest to the Awarding Committee a range of marks within which they expect the grade boundary to be set.

The Awarding Committee comprises senior examiners who have been involved in the writing of the question papers, marking of scripts and the setting of standards in previous years (except in the case of a new specification). The aim of the committee is to determine the grade boundaries for each externally assessed unit (which is usually one examination question paper) in line with the standards of previous years and across different specifications. Thus the central concern of the awarding process is not one of score ***interpretation*** but one of score ***equating***; the committee's aim is not to ***set*** a standard but to ***maintain*** one.

During the Awarding Meeting the Committee start by looking at scripts at the bottom of the range (for say grade A) and note the mark at which it is unclear that the scripts no longer have the unique characteristics of a grade B. Then the Awarders start reading scripts at the top of the range and record when it is unclear that the script is worthy of a grade A. The resulting range of marks is known as the zone of uncertainty. It should be noted that when examiners study the scripts they also refer to archive scripts at the appropriate boundary and to grade descriptors. Once the zone of uncertainty has been determined the examiners refer to statistics to choose the appropriate grade boundary. This is repeated for each unit for all judgementally awarded grades. The remaining grade boundaries are determined arithmetically. The procedures for Awarding are given in full in Qualifications and Curriculum Authority (QCA) (2003/4).

3 Criterion and norm referencing

Authors often divide educational assessment systems into criterion referencing and norm referencing, for example: *The notion of a standard in an assessment system is defined as the attachment of specific meanings to specific test or examination scores. When these meanings are in terms of the performance of a group of individuals, the standard can be described as norm referenced (when the individual is not a member of the reference group), and when the meanings are in terms of explicit criteria, the standard is described as criterion referenced*. (Wiliam, 1996b, 293).

Assessment measures in a criterion referenced system are supposed to be based on an absolute measure of quality, whilst assessment measures in a norm referenced system are supposed to be based on a relative standard. Criterion referencing should involve the use of detailed criteria which candidates must meet to pass or gain a given score[iii]. The criteria are intended to be such that an observer can determine whether the candidate's performance has met the criteria.

Assessments that are criterion referenced indicate the performance of a candidate in terms of a criterion standard which provides information as to the degree of competence of an individual student, independent of the performance of others. For assessments which are norm referenced an individual candidate's score is compared to the achievement of a standard reference group. It is on the basis of the candidates' achievement and where they are ranked in relation to the distribution of achievement of the standard reference group, that grades/passes are awarded. Note that the candidates are not members of the reference group.

Some educational measurement specialists might also list domain referencing as the third way of interpreting the meaning of an assessment score. Domain referencing involves specifying a domain so that an item bank of all the possible questions for a particular domain can be generated. Tests are made by randomly choosing a sample of questions. The meaning

of a score is that the percentage of questions that the candidate got correct is the percentage of the domain that they know. The limitations of domain referencing are that tests would include both good and bad questions, as it is impossible to write all the possible questions in any particular domain. But some authors consider criterion referencing (which is defined in many different ways) to be conceptualised in two different ways: (1) as domain referencing and (2) as derived from mastery learning theory. Berk (1980, 5) quotes Popham (1978, 93): *A criterion referenced test is used to ascertain an individual's status with respect to a well defined behavioural domain.* Then Berk (1980, 5) adds that: *Such tests are often referred to as domain referenced….An alternative conceptualisation of criterion-referenced measurement derived from mastery learning theory….It is used to classify students as masters and nonmasters of an objective in order to expediate individualised instruction.*

It is beyond the scope of this paper to consider the differences between criterion and norm referencing along with their advantages and limitations. But they are well rehearsed by many authors, for example, Glaser (1963), Linn (1993), Gipps (1994) and Wiliam (1996a &b).

Each of these forms of referencing is concerned with how a score might be interpreted, or given meaning. Obviously criterion, domain and norm referencing are situated in a social context which also affects the meaning of the score. For example, the meaning of a score of 50% will be different in 1950 and 2000; or in English Literature and forklift truck driving, irrespective of whether the score is criterion, domain or norm referenced. Criterion, domain and norm referencing are also all abstract 'pure' descriptions of referencing systems which can rarely function in practice in their pure form.

Linn(1993, 140) states that: *It is generally recognised now that the terms norm referenced and criterion referenced, as applied to instruments, are somewhat misleading. It is the interpretation given to scores, not the instruments themselves, which are criterion-referenced or norm-referenced*.

Until the late 1960s the General Certificate of Education public examinations were usually said to be norm referenced. It is generally accepted that public examinations in England and Wales (indeed many public examinations) are neither criterion referenced nor norm referenced, rather their scoring is derived from both. The process of determining grade/level thresholds for UK public examinations has been described by different authors as cohort referencing, limen referencing and weak criterion referencing. Construct referencing has been described as a way of assigning grades to GCSE coursework. Arguably none of these systems are reference systems (ways of assigning meaning to a score) they are methods of deciding which mark should be the grade boundary[iv]. In this paper the Awarding process and associated concepts are revisited with the aim of scrutinising the different terminology that has been used to describe Awarding and to consider how it might be best conceptualised.

4 Cohort referencing

Cohort referencing is ensuring that the proportion of candidates awarded each grade remains the same from one examination session to the next. The grade boundaries are therefore fixed so that a predetermined proportion of candidates achieves each grade. Note that the candidates are part of the cohort group (Wiliam, 1996a, Baird et al. 2000). This was the original GCE A level grading system and is now the basis of much of the statistical analysis used in grading GCSEs and GCEs (Baird et al, 2000). However, as noted above, both statistics and professional judgement are used to set GCSE and GCE standards. The advantage of cohort referencing over norm referencing is that the scores on the test and the grades awarded do not depend upon the reference group, which might be unrepresentative, inappropriate or out of date.

4.1 What are the limitations of cohort referencing?

A limitation of cohort referencing is that it only indicates the ranking of a candidate in relation to other candidates who took the examination in the same session. That is, the

standard is determined by the cohort rather than carried over from previous years. There might be differences between various cohorts and so the quality of the performance of an average candidate in, say, 2001 might be better than that of an average candidate in 1996, as the test has become more familiar. Obviously, expectations about candidates' performance change over time. Comparability studies[v] and Awarding are arguably biased towards the past in the sense that they aim to ensure that standards from previous years have been maintained. Therefore, there needs to be a check on the quality of the achievement and what has been achieved, as well as the quantity of candidates achieving particular scores or grades.

4.2 How well does cohort referencing describe the Awarding process?

Cohort referencing can only be said to describe part of the Awarding process i.e. the statistics that are used as indicators of where to put grade boundaries.

5 Limen referencing

This paper will focus upon limen referencing, about which there is little published literature. Described succinctly, limen referencing is the comparison of a particular candidate's performance with consensus standards (French et al, 1988). The consensus standards they refer to are the consensus standards as perceived by senior examiners, these views of standards are carried forward from the previous examination session.

Christie and Forrest (1982) were the first to define *limen referencing* a concept which they borrowed from information theory. They explain that the model of subject competence which characterises public examinations in England and Wales is quite different to the models of subject competence in the research literature.

There is a clearly enunciated achievement domain. There are cutting scores though these are always modified on the basis of the scrutiny of candidates' responses rather than fixed in advance by subject matter. The grades have at least the force of custom and the chief examiners can identify borderline performances with some accuracy. The grades do not carry with them any explicit definition of the achievements they imply. We shall dub it a limen referenced model to emphasise it's role in dealing with fuzzy signals implying an interdependence between transmitter and receiver. Christie and Forrest (1982, 57).

Limen referencing was not taken further than the level of a loose framework of customary practice to the status of a descriptive model, it was certainly not taken to the level of a working or prescriptive model (Christie and Forrest, 1982). For this to happen they explained that there would need to be (1) explicit subject related criteria associated with each grade and (2) a mathematical model to explore the relationship of multiple achievement criteria to grade boundaries. Christie and Forrest (1982) cited French (1980), Fishburn (1967) and Krantz et al (1971, chapter 9) as offering suitable mathematical models.

The subject related criteria planned for the GCSE were: *a much softer notion than that of strict criterion referencing, implying, we understand, a greater role for the examiners' judgement and interpretation* (French et al. 1988, 22). The attempts to write grade criteria for GCSE failed for a number of reasons. For example, they were too complex to be useful, and criteria do not accord with a compensation approach[vi]. For further details about grade criteria see Gipps (1990) and Cresswell and Houston (1991). French et al (1988) envisaged that grade descriptors rather than grade criteria would have a role to play in Awarding, where descriptors are intended to make standards clearer to the non-examining public and increase the degree of consensus amongst examiners.

French et al (1987b) use the term limen referencing to describe the judgmental ordering which examiners carry out during grading. Limen is a concept which has been commonly used in psychological work on perception - as in *subliminal*. Limen means a threshold and subliminal means below the threshold; below a certain stimulus intensity there is no sensation and this is the limen or threshold where the stimulus is assumed to be zero. French et al (1987b) use the term limen in the sense that an A grade boundary is the limen (threshold) at which the unique qualities of A grade achievement as perceived by the senior

examiners as *detectable*, and they are *undetectable* below the grade boundary. The same principle applies to other judgementally awarded grades. The notion that there are unique characteristics which examiners use to distinguish between adjacent judgementally awarded grades was suggested by Pollitt and Murray (1996).

The notion of a limen, threshold or grade boundary is not a ground breaking idea. What is important is the realisation that standards are not objective measures, they are the subjective judgements, or the consensus perception of senior examiners, or experts (French et al. 1987a). This was also later argued by Cresswell (2000). French et al. (1988, 17) argued that: *The quality of the candidate's performance does not exist within the candidate or within his/her work. It is a mental construct of an observer of that performance: it exists within the perception of the examiner. ....It is formed gradually and undergoes many revisions as the examiner studies the candidate's script. We believe that the purpose of many of the quantitative procedures used in examination assessment is to help the examiner form his perception and his judgement of the candidates' performances in a manner that is fair and consistent to all candidates*. Although the last quote is in the singular, French et al. (1988) acknowledge that there is more than one examiner making the judgements, they say: *Christie and Forest term the comparison of a particular candidate's performance with these consensus standards limen referencing* (French et al, 1988, 23).

French et al. (1988) also describe how limen referencing is operationalised. The examiners have in mind a typical candidate(s) whose performance represents the standard and quality for a particular grade. Whilst grading, the examiners will make comparisons between the ranked performance of examined candidates and their perceptions of grade-typical performances. The limen referencing decisions are not made in a vacuum, they are supported by:-

- designing the question papers with particular standards and grade thresholds in mind;
- reading archive scripts;
- cohort referencing.

Afterwards, with all the candidates provisionally graded, the examiners engage in an activity French et al (1988) call *borderlining*. The examiners consider carefully, individual candidates who might have been misgraded or misjudged: *In terms of limen referencing this means that candidates' scripts are reread and their performances compared directly with the consensus standards* French et al (1988, 23).

5.1 What are the limitations of limen referencing?

It is beyond the scope of this paper to consider in depth the limitations of qualitative judgements in Awarding (and limen referencing in the examining context) as there is already a good deal of research literature about this topic, so references to important works will suffice. Wiliam (1996b) points out that the maintenance of standards requires that the standard setters (in the case of GCSE and GCE senior examiners and accountable officers) must be full participants in a community of practice and that they must be trusted by the users of assessment results. Obviously if this trust has been lost the validity of the currency of qualifications and associated grades falls apart.

Good and Cresswell (1988) and later Cresswell (2000) found that Awarding Committee decisions moved grade boundaries in the appropriate direction, but their judgements of how large the adjustment should be did not match statistical predictions. The author's experience of attending a variety of Awarding Meetings in different subjects over the past 5 years suggests that the former research is familiar to Awarders who make decisions about comparable grade boundaries for different tiers. Cresswell (2000) states that: …*qualitative judgement alone is inadequate as a method of maintaining examination grade standards from year to year because it does not take sufficient account of changes in the difficulty of successive year's examinations* (Cresswell, 2000, 94). Cresswell (2000) adds that Murphy et al. (1996) and Cresswell (1997) outline the psychological and sociological reasons for the limitations of qualitative judgements as a way of maintaining standards. There are of course other factors which affect Awarders' judgements, for example, Scharaschkin and Baird (2000)

found that Biology examiners were more likely to consider inconsistent[vii] scripts as worthy of lower grades than consistent scripts or scripts of average consistency. Sociology examiners were more likely to classify consistent scripts as worthy of higher grades than scripts of average consistency.

Although it is not mentioned specifically by Christie and Forrest (1982), or French et al's (1988) descriptions of limen referencing, it is reasonable to assume that the examiners' perceptions of standards have developed historically.

Wolf (1995) explains that it has been repeatedly found in assessment that assessors can rank candidates' work consistently with one another but that they find it difficult to judge whether a candidate's work meets a given standard or criteria. Therefore assessors would find it difficult to undertake limen referencing in the strict sense of comparing a script against a standard. However, they would find it easier to judge whether a piece of work was of a better or worse standard than other work. So when limen referencing is taken to include the comparison of live scripts with archive scripts it becomes a practical methodology.

5.2 How well does limen referencing describe the Awarding process?

When examiners are finding the zone of uncertainty, the limen referencing principle of identifying when the signal becomes detectable, is applied. But it is applied in what some might consider to be an upside down approach (as the Awarders are looking for a threshold above which they cannot detect the unique qualities of grade B).

The Awarders agree upon a range of marks (zone of uncertainty) rather than an individual mark or limen as this allows for:-

- the possibility that examiners cannot detect differences in the quality of whole examination scripts to a single mark (although they can make distinctions between the quality of work which was awarded adjacent marks for a given question);
- some minor disagreement amongst individuals;
- limen referencing to be used with cohort referencing.

The other aspects of grading to which limen referencing literature refers are the use of statistics (cohort referencing) and the examiners who wrote the question papers and mark schemes.

The current Awarding practice as defined in the Code of Practice does not include borderlining as described by French et al (1988). However, there is a process called marking review to ensure that, in cases where there remains doubt about whether the marks given to a candidate or group of candidates are acceptable, action is taken where necessary. The scripts of such candidates will be remarked by senior examiners and the new mark will replace the old mark and their overall mark and grade will be recalculated (QCA, 2003/4).

In summary, limen referencing (as described by French et al. and Christie and Forrest) is part of current grading procedures but it is used with other procedures to grade GCSEs and GCEs.

6 Construct referencing

Wiliam (1996b) states that there is evidence (some anecdotal and some unpublished) that teachers from a 100% coursework GCSE in English quickly learnt[viii] to agree what grade an example of work was worth. The teachers did not necessarily agree on which aspects of the work were most significant in making the work worthy of a particular grade. Wiliam (1998) adds that there was no attempt to prescribe learning outcomes. *The touchstone for distinguishing between criterion- and construct-referenced assessment is the relationship between the written descriptors (if they exist at all) and the domains. Where written statements collectively define the level of performance required (or more precisely where they define the justifiable inferences), then the assessment is criterion referenced. However where such statements exemplify the kinds of inferences that are warranted, then the assessment is, to an extent at least, construct referenced* (Wiliam, 1998, 6)

Obviously a level or a grade has different meanings for different individual teachers, candidates and other stakeholders. But these different meanings of the level or grade are not

necessarily mutually exclusive. Wiliam (1996b) argues that the teachers had internalised a notion of 'levelness'. It is assumed that 'levelness' is the same as grade worthiness (as the teachers could agree on a grade). He adds that: *....the consistency in teachers' judgement (where it exists) arises out of 'levelness' in English; in other words, the assessment is 'construct referenced'. It could be argued that construct referenced assessment is just a form of norm referenced assessment, where the reference group are the national cohort of students, but there is an important difference in emphasis. In norm-referenced assessment the standard exists in the performance of the reference group. In construct referenced assessment, whatever standard that exists does so in the heads of the teachers involved in the assessment*. (Wiliam, 1996b, 298).

One might ask whether the construct changes for the different grades/levels. Wiliam (1996b) does not tackle this question. From work about grade/level descriptors e.g. Pollitt and Murray (1996) it can be assumed that there are distinguishing features of a level/grade which teachers look for when they are assessing. The distinguishing features are the extras that people who achieve that level (grade) can do in addition to being able to do what other candidates who achieved lower levels (grades) can do. In this situation the construct has two elements captured in Wiliam's phrase *levelness in English*. The first element is what can be awarded marks for that test or examination (in this case GCSE English) and what distinguishes performance at different levels or grades (levelness).

The descriptions given by Wiliam (1996b; 1998) suggest that construct referenced assessment is used when referring to marking and/or assigning grades but not to choosing grade boundaries.

6.1 What are the limitations of construct referencing?

Construct referencing is concerned with qualitative judgements. The problems with the qualitative judgements made by Awarders have been outlined above 5.1.

Arguably one limitation of construct referencing is that it is not objective. Wiliam (1996b, 298) says that: *In construct referenced assessment, objectivity is achieved .....through intersubjectivity*.

6.2 How well does construct referencing describe the Awarding process?

Construct referencing describes the qualitative judgements made by teachers about assigning grades or levels. So far it seems that this term has not been used to refer to the choosing of grade boundaries. The construct or the general quality of work required for the grade is illustrated in the question paper, grade descriptors, mark scheme, archive scripts etc. In this sense the standards are not just in the heads of examiners. There is some interaction between the list of tools given above that indicates the standard and the examiners' views. In Awarding examiners use their constructs of two adjacent grades to find the zone of uncertainty which is part of the Awarding process. Such qualitative judgements, combined with cohort referencing, describe Awarding. Baird et al. (2000) explain that it is not just senior examiners who set standards (as suggested by construct referencing), Awarding Body statisticians and Chief Executives are also involved in the process. The responsibilities of the accountable officer and Awarders are described by QCA (2003/4).

7 Weak Criterion referencing

Later, Baird et al (2000) developed the concept of weak criterion referencing (WCR) which is defined as relaxing the notion of criterion referencing. It involves: *…maintaining the general quality of examination performance required for each grade, given the difficulty of the examination but not demanding evidence of specific achievements* (Baird et al. 2000, 215). It is a judgement task undertaken by senior examiners in the setting and maintaining of GCSE and GCE standards. In making their judgements they take into account the demands of the examination question(s) and/or tasks and the context of the assessment which changes from year to year. To do this the Awarders must mentally aggregate the candidates' answers

to a number of questions and compare sets of answers from different examination sessions. If specifications have changed substantially over time then the challenge of making comparisons increases.

Baird et al. (2000, 216) say that: *The standard setting judgement is essentially a multi-attribute decision making task, in which awarders are expected to integrate the information they have about candidates' performances from a range of questions within examination papers. The fundamental requirement upon awarders is that they are capable of detecting the scores on different papers at which identical levels of performance are observed and to do this they must be able to allow for the different levels of question difficulty*[ix]. The WCR approach assumes that Awarders will be able to make judgements about the standards of candidates' performances taking different contexts, question difficulty and varied levels of consistent performance into account. However, if Awarders could judge and adjust question difficulty then grade boundaries could be the same from year to year and there would be no need for Awarding Committees (Baird et al. 2000).

The WCR approach might be used when assessments use a compensation approach to giving credit. The WCR approach assumes that Awarders are able to judge consistent and inconsistent performances resulting in the same examination score with the same level of leniency. Additionally WCR does not specify *precisely* what grades are awarded for, although more general statements like grade descriptors can be written. For examples of grade descriptors, their limitations and a method of writing them, see Greatorex et al (2001).

7.1 What are the limitations of weak criterion referencing?

Generally experts are poor at making accurate predictions (Camerer & Johnson, 1991). This phenomenon emerges in examining in a number of ways. For example, Kingdon and Stobart (1988) explained that examiners found it difficult to target or pitch questions for a particular ability range for tiered examinations (some GCSEs are tiered). For weak criterion referencing to be effective, examiners must be able to judge the difficulty of the questions.

The limitations of qualitative judgements made by examiners are listed above in 5.1 and apply equally to limen referencing and weak criterion referencing. Further difficulties are explained above, namely that Awarders judgements are affected by the consistency of candidates' performances (Scharaschkin and Baird, 2000). There is also evidence from Good and Cresswell (1988) and Cresswell (2000) that Awarders cannot assess the full impact of question difficulty on candidates' performance.

The work by Baird et al (2000) implies that WCR is used within the same specification or subject, and that it cannot be applied across subjects. Additionally, work by Newton (1997) suggests that weak criterion referencing can only be used over relatively short time periods. If there is a long time period between when the examinations are taken, say 20 years, then comparisons are made about relative attainment in context, as the standards need to be understood as relative to societal expectations.

7.2 How well does weak criterion referencing describe the Awarding process?

WCR as defined by Baird et al (2000) summaries the Awarding process but, the difficulty with WCR is that it is defined as criterion referencing without precise written criteria. Without precise written criteria a criterion referenced system is no longer criterion referenced. It is possible that the 'weak' in WCR refers to the lack of precision in defining the standard. The 'criteria' are the unarticulated quality of the examination performance required. This general quality might be illustrated by archive scripts, grade descriptors, mark scheme and/or the examiners' tacit knowledge of the quality of work expected. The definition of WCR fits with the process of Awarding but the name WCR is arguably misleading.

8 How might Awarding be conceptualised?

As yet literature searches have not revealed why the Awarding process was first undertaken in the manner described above. Whilst the practice appears sensible and systematic it does not have a conceptual/theoretical basis. Consequently terms e.g. limen

referencing/WCR have been developed to describe the practice rather than the practice being developed from first principles. If this were to be undertaken it must be acknowledged that it is impossible to develop Awarding procedures independently of the design of the qualification or examination system.

If the above are limited for describing and understanding Awarding should it be considered from a different perspective or should an alternative be suggested? Pollitt and Elliott (2003 a & b) have suggested that Awarding as described above could be replaced by a large scale Thurstone paired comparison (described below). This suggestion currently exists as an idea rather than as a planned operational procedure where all the loose ends have been tied up. The scripts from all candidates would be ranked, based upon internal or external assessment. This process would be standardised[x] (for external assessment[xi]) and moderated[xii] and standardised (for internal assessment[xiii]).

Examiners, teachers and/or other stakeholders would compare pairs of scripts[xiv] from the current examination session and previous examination session(s) from the same (or most similar) syllabus. For each comparison participants would individually scrutinise the candidates' work for a short amount of time and decide which is the better work. The results of these many decisions would be analysed to form a consensus scale. The analysis would indicate where the grade boundaries should be, based upon previous grade boundaries by including some archive scripts in the paired comparison procedure. Following this suggestion, the standard from previous examination sessions would be carried forward to the live examination session. As grade boundaries are set at the unit level and then aggregated to give a qualification level grade boundary, it might be possible to grade using this method at the unit or qualification level. Indeed if this concept is taken to it's extreme, marking could be eradicated, scripts would just need to be ranked. Additionally it relies on judges ranking rather than comparing students' work against a standard, it is reported consistently in the research literature that judges can successfully do the former but not the latter (Wolf, 1995).

The problem with omitting marking from the process is that candidates/stakeholders might not trust a system that only scrutinises each script for a short length of time. The process follows the definition of WCR - maintaining standards - although it does not easily accord with the notion of criterion referencing. It overcomes the problems which can occur with qualitative judgements made by Awarders, for example, knowing which way the boundary should be moved but not by how far. However, it still requires the Awarders to be able to predict or estimate the difficulty of the questions and/or examination and take this into account. A further problem with this approach is that it is obvious which is the live examination scripts and which are the archive examination scripts. This might be a source of bias in the judgements made in the process. It is possible to detect bias statistically in a Rasch analysis but whether substantive bias exists, is a matter of professional judgement by an analyst.

If electronic marking and the item level data were available, then statistics indicating how well the individual questions had worked (in terms of reliability[xv], difficulty and discrimination[xvi]) could be used to decide where would be an appropriate "grade boundary" on each question. These "grade boundaries" would be aggregated to give a grade boundary for the paper. This is essentially the Angoff procedure[xvii] made empirical. The other possibility is to sort questions in order of difficulty and then decide which is the most difficult question candidates must complete correctly to gain a particular grade. This is known as bookmarking[xviii]. One advantage of these approaches is that the senior examiners will not be expected to predict or estimate the difficulty of the questions or examination paper, which they do not do well. The disadvantage of both approaches is that they do not allow for compensation.

Another approach might be to use statistics alone (rather than the Principal Examiners' judgements and statistics) to predict the range of marks which were likely to be appropriate as a grade boundary. Awarders would then choose a grade boundary from within that range, based upon their notions of grade worthiness. This would overcome the potential problem[xix] of Awarders knowing which way to move grade boundaries but not knowing how far they should be moved.

9 Conclusions

The best description of Awarding is a combination of cohort referencing and professional judgements made by senior examiners and accountable officers. The definitions or descriptions of limen referencing and WCR describe Awarding well but the terms do not describe the process well.

The views of Awarding as described above perpetuate the principle that standards should be maintained rather than providing an interpretation of the outcomes. Perhaps the emphasis should be on improving the standard of performance for all candidates, thus creating modern valid assessments and interpreting (given meaning or referencing) outcomes to indicate to stakeholders what students know and can do rather than maintaining standards. Obviously this is difficult to achieve and creates it's own problems, for example, descriptions of what candidates can do are interpretable and changes in education systems create additional work for teachers and pupils, and make systems more difficult to understand (simply through lack of familiarity). Such a system would need to begin by identifying what stakeholders would like candidates to know and do, and for assessment professionals to consider what are valid and fair methods of assessing and grading the qualifications. How the outcomes might be interpreted (given meaning or referenced) in a way that has meaning to all stakeholders and accommodates the multiple purposes of public examinations could then be developed from first principles.

10 Bibliography

Baird, J. Cresswell, M. and Newton P. (2000). Would the real gold standard please step forward? *Research Papers in Education*, 15, (2), 213-229.

Bell, J. F. and Greatorex, J. (2000). *A Review of Research in Levels, Profiles and Comparability*, A report to the Qualifications and Curriculum Authority.

Berk, R. A. (1980). (Editor) *Criterion-Referenced Measurement The State of the Art,* The John Hopkins University Press: Baltimore.

Camerer, C. F., & Johnson, E. J. (1991). *The process-performance paradox in expert judgment: How can experts know so much and predict so badly?* In K. A. Ericsson & J. Smith (Editors.), Towards a general theory of expertise: Prospects and limits (pages 195-217). Cambridge Press: New York.

Christie, T. and Forrest, G. M. (1982). *Defining Examination Standards*, Schools Council Research Studies: London.

Cizek, G. J. (2001). *Setting Performance Standards: Concepts, Methods and Perspectives*. NJ: Lawrence Erlbaum Associates: Mahwah.

Cresswell, M. (2000). *The Role of Public Examinations in Defining and Monitoring standards*, In H. Goldstein and A. Heath, Educational Standards, (pages 69 to 120). Oxford University Press: Oxford.

Cresswell, M. (1997). *Examining Judgements: Theory and Practice of Awarding Public examination grades* . PhD thesis, University of London Institute of Education: London.

Cresswell, M. and Houston, J., (1991), Assessment of the National Curriculum - some fundamental considerations, *Educational Review*, 43, 1, 63-78.

Fishburn, P. C. (1967). Methods for estimating additive utilities, *Management Science*, 13, 435-453.

French, S. (1980). *Measurement theory and Examinations* (Notes in Decision Theory, Note No 88) Department of Decision Theory, University of Manchester.

French, S., Willmott, A. S. and Slater, .J. B. (1987a). *Decision Analytic Aids to Examining*, School Examinations and Assessment Council: London.

French, S., Slater, J. B., Vassiloglou M. and Willmott A. S. (1987b). *Descriptive and Normative Techniques in Examination Assessment,* Occasional Publication, University of Oxford Delegacy of Local Examinations. OIASL: Oxford.

French, S., Slater, J. B., Vassiloglou, M., and Willmott, A. S. (1988). *The role of Descriptive and Normative Techniques in Examination Assessment*, In Black, H. D. and Dockrell, B. (Editors) (1988). Monograph of Evaluation and Assessment Series No. 3, Scottish Academic Press: Edinburgh.

Gipps, C. (1990). *Assessment a teachers' guide to the issues*, Hodder and Stoughton: London.

Gipps, C. (1994). *Beyond testing: Towards a theory of educational assessment*, Falmer Press: London.

Glaser R (1963). Instructional technology and the measurement of learning outcomes: Some questions, *American Psychologist*, 18, 519-21.

Good, E. F. & Cresswell, M. J. (1988). Grade awarding judgements in differential examinations, *British Educational Research Journal*, 14, (3), 263-281.

Greatorex, J. (2003). *'Examination and assessment in Curriculum 2000.'* In L. Le Versha, and G. Nicholls, (2003) *Teaching at Post-16 Effective Teaching in the A-level, AS and VCE Curriculum* (pages 16-24), Kogan Page: London.

Greatorex, J., Elliott, G. and Bell, J., (2002). *A Comparability Study in GCE AS Chemistry Including parts of the Scottish Higher Grade Examinations*, A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2001 Examination and organised by the Research and Evaluation Division of the University of Cambridge Local Examinations Syndicate on behalf of the JCGQ.

Greatorex, J., Johnson, C. & Frame, K. (2001). Making the grade - developing grade profiles for accounting using a discriminator model of performance. *Westminster Studies in Education*, 24, (2), 167-181.

Kingdon, M. and Stobart, G. (1988). *GCSE Examined*. The Falmer Press: London.

Krantz, D. H., Luce, R. D., Suppes, P. and Tversky, A. (1971*). Foundations of Measurement Volume 1*. New York: Academic Press.

Lewis, D. M., Mitzel, H. C. & Green, D. R. (1996). *Standard setting: A Bookmark approach*. In D. R. Green, (Chair), IRT based standard setting procedures utilising behavioural anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

Linn, R. L., (1993). (Editor). *Educational Measurement*, Oryx Press: Phoenix, USA.

Mitzel, H. C., Lewis, D. M., Patz, R. J. and Green D. R. (2001). *The Bookmark procedure: psychological perspectives*. Chapter 9 in G. J. Cizek (Editor) Setting Performance Standards: Concepts, Methods and Perspectives. Lawrence Erlbaum Associates: Mahwah: NJ.

Morrison, H., Busch, J. and D'Arcy, J. (1994). Setting Reliable National Curriculum Standards: A guide to the Angoff Procedure, *Assessment in Education*, 1, 181-199.

Murphy R. J. L., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. and Gower, R. (1996). *The Dynamics of GCSE Awarding*: Report of a Project conducted for the School Curriculum and Assessment Authority: London.

Newton, P. (1997). Examining Standards over time, *Research Papers in Education*, 12, (3), 227-248.

Nuttall, D. L. and Willmott, A. S. (1972). *British Examinations Techniques of Analysis*, National Foundation for Educational Research: London.

Pollitt, A. & Elliott, G. (2003a). *Monitoring and investigating comparability: a proper role for human judgement*. Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', April 3rd-4th. http://www.ucles-red.cam.ac.uk/conferencepapers/QCA2003APGE1.pdf

Pollitt, A. & Elliott, G. (2003b). *Finding a proper role for human judgement in the examination system*. Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', April3rd-4th. http://www.ucles-red.cam.ac.uk/conferencepapers/QCA2003APGE02.pdf

Pollitt, A. and Murray, N. L. (1996). *What raters really pay attention to*, in M. Milanovic and N. Saville (1996) (Editors) Studies in language Testing: 3, performance testing, cognition and assessment. Cambridge University Press: Cambridge.

Popham, W. J. (1978*). Criterion-referenced measurement*. Prentice Hall: Engelwood Cliffs, N. J.

Qualifications and Curriculum Authority, (2003/4). *GCSE, GCSE in vocational subjects, GCE, VCE, GNVQ and AEA Code of Practice 2003/4*, Qualifications and Curriculum Authority: London.

Sadler D. R. , (1987). Specifying and Promulgating Achievement Standards, *Oxford Review of Education*, 13, (2), 191-209.

Scharaschkin, A. & Baird, J. (2000). The effects of consistency of performance on A level examiners' judgements of standards, *British Educational Research Journal*, 26, (3), 343-357.

Wiliam, D. (1996a). Meanings and consequences in standard setting, *Assessment in Education*, 3, (3), 287-307

Wiliam, D. (1996b). Standards in examinations: a matter of trust? *The Curriculum Journal*, 7, (3), 293-306.

Wiliam, D. (1998). *Construct-referenced assessment of authentic tasks: alternatives to norms and criteria*, Paper presented at the 24th Annual Conference of the International Association

for Educational Assessment - Testing and Evaluation: Confronting the Challenges of Rapid Social Change, Barbados, May 1998.

Wolf, A. (1995). *Competence Based Assessment*. Open University Press: Buckingham.

---

[i] GCEs (A Levels) are general subjects normally taken by 18 year olds. They are generally a selection mechanism for higher education.

[ii] GCSEs are normally taken by 16 year olds in England and Wales. They are also taken as school examinations in different countries around the world. In England and Wales they are often a prerequisite for A level study. The first GCSEs were taken in 1988. They replaced O levels and Certificates of Secondary Education (CSEs).

[iii] For the purposes of this paper score is taken to refer to both GCSE/A level grades and/or final marks. GCSE/A level grades are determined by marks and therefore grades and marks are in the same order.

[iv] A grade boundary is the minimum mark required to be awarded a particular grade.

[v] Comparability studies are research exercises to identify whether the standard of one examination or qualification is broadly comparable to (same as) another. For more information about comparability see Bell and Greatorex (2000), for an example of a comparability study see Greatorex et al. (2002).

[vi] That is candidates can make up for their weaknesses by gaining marks for their strengths. Candidates can reach the same grade through different routes as they have different strengths and weaknesses. In this model of Awarding, grades are awarded on the aggregation of marks. In contrast some assessments use a ***conjunctive model*** which is an evaluation or scoring procedure that requires the candidate to attain a minimal level of performance on all attributes assessed, for example, this was the situation with GNVQs. Another alternative is a ***disjunctive model*** when an evaluation or scoring procedure requires the candidate to achieve a minimal level of performance on only one of the attributes assessed.

[vii] Scharaschkin and Baird (2000) were considering the consistency of performance of candidates in scripts in terms of the proportion of marks that candidates achieved for each question.

[viii] The teachers were trained and given feedback from an expert on their marking (Wiliam, 1998).

[ix] Difficulty is measured by calculating the mean mark achieved by candidates on that question (as a proportion of the total marks available) or examination paper. There are also more sophisticated measures of difficulty.

[x] *Standardisation A process relating to both internal and external assessment, by which the Awarding Body ensures that the mark scheme or assessment criteria for a unit or component are applied consistently by examiners or moderators. For example, the process may include a meeting of examiners or moderators to consider the mark scheme or assessment criteria in detail* (QCA, 2003/4, 57*). Internal standardisation A process carried out by centres in relation to internally-assessed work to ensure, for a particular specification, that all candidates are judged against the same standards, across different assessors and teaching groups*. (QCA, 2003/4, 56)

[xi] *External assessment A form of assessment in which question papers and tasks are set by the Awarding Body, taken under specified conditions (including details of supervision and duration) and assessed by the Awarding Body. This includes Awarding Body set assignments.* (QCA, 2003, 56)

[xii] *Moderation The process through which internal assessment is monitored by the Awarding Body to ensure that it is reliable, fair and consistent with required standards.* (QCA, 2003, 56). Internal assessments are all moderated. This can involve postal moderation and/or a moderator visiting the centre, depending upon the nature of the coursework. For A level coursework and VCE portfolios centres are asked to ensure that the rank ordering of the internally assessed units is correct. The role of a moderator is to ensure that a mark awarded for work of a given standard is the same from centre to centre. Moderators do not award a grade Greatorex (2003).

[xiii] Internal assessment A form of assessment that does not meet the definition of external assessment. (QCA, 2003). Internal assessment is generally assessed by the candidates' teacher/tutor. In the case of GCEs internal assessment is often called coursework. (Greatorex, 2003)

[xiv] "Script" could refer to an examination script or to project work etc. It might refer to a candidates work at the unit (question paper) level or to the entire work of a candidate that is to be assessed for the qualification.

[xv] Reliability is the consistency with which a test(s) measures a particular trait. The internal consistency reliability can be measured at the question paper level using Cronbach's alpha or Backhouse's P when there is question choice. Internal consistency reliability can be interpreted as the correlation between a test and all other possible tests which might be constructed from a hypothetical universe of questions measuring the same trait. The composite reliability of a question paper can be

measured using a formula by Nutall and Willmott (1972). All these reliability coefficients are a form of correlation coefficient.

[xvi] Discrimination is about whether a question distinguishes effectively between candidates of different calibre. Product moment correlations between marks at the question and paper level illustrate whether candidates who do well on a question do well overall. Similarly the mark on each question can be correlated with the marks on the rest of the paper or the mark on the question can be correlated with a mark or test score on another examination paper or another criterion e.g. the whole qualification or something else.

[xvii] For further information about the Angoff procedure see Morrison et al (1994) and Cizek (2001).

[xviii] For more information about bookmarking see Lewis et al (1996) and Mitzel et al (2001).

[xix] Cresswell (1997) gave evidence for this problem which is discussed earlier.