CAMBRIDGE ASSESSMENT

Contemporary GCSE and A-level Awarding: A psychological perspective on the decision-making process used to judge the quality of candidates' work.

A paper presented at BERA 2007.

Jackie Greatorex
Research Division
Cambridge Assessment

Address for correspondence
Jackie Greatorex
Research Division
Assessment Research and Development
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG
UK
E-mail: greatorex.j@cambridgeassessment.org.uk
Telephone: 44 (0)1223 553835
Fax: 44 (0)1223 552700

www.cambridgeassessment.org.uk

UNIVERSITY of CAMBRIDGE
Local Examinations Syndicate

Cambridge Assessment

Cambridge Assessment is the brand name of the University of Cambridge Local Examinations Syndicate, a department of the University of Cambridge. Cambridge Assessment is a not-for-profit organisation.

Abstract

General Certificate of Education and General Certificate of Secondary Education are high stakes qualifications and the results are important for candidates' future education and employment. At the moment many aspects of examining in England are being modernised, e.g. the introduction of e-marking in some contexts. As part of this modernisation process there is renewed interest in innovative methods of recommending grade boundaries. This paper is intended to contribute to the current public debate about appropriate methods for recommending grade boundaries. The purpose of this paper is also to address the question of which Awarding procedure successfully avoids the limitations and plays on the strengths of human judgement. To tackle this question four methods of recommending grade boundaries are considered: 'current Awarding', Thurstone pairs, 'remote Awarding' and rank ordering. Research evidence is then reviewed to evaluate each of these methods against common criteria.

At the moment only 'current Awarding' is used to recommend grade boundaries in non-experimental settings, according to the research literature in the public domain. The other methods of recommending grade boundaries are methods which might be adopted in the future, and are currently at the trialling stage. Current Awarding practice meets the regulatory criteria set by the Qualifications and Curriculum Authority. However, it is concluded that whilst a body of knowledge is developing further research is needed before we can evaluate how well the four methods measure up to each criterion. Consequently, current remarks about which is the best method of Awarding might not be entirely research evidence based. The review proposes that methods like rank ordering and Thurstone pairs have some advantages over current or remote Awarding. Research is needed in the area of judges' cognition and the construct validity of some of the methods. Consequently, there is ongoing research about these issues. The intention of the work in progress is to contribute to the public debate and to inform future decisions about which method would be the most effective.

1. Introduction

General Certificate of Education (GCE)[i] or General Certificate of Secondary Education (GCSE)[ii] grades affect thousands of candidates many of whom are hoping to go on to university or further study. Therefore, it is important that Awarding procedures (in which grade boundaries[iii] are recommended) result in an appropriate grade for each candidate. The GCSE and GCE A-level Awarding procedures rest on a number of human judgements about the difficulty[iv] and demand[v] of current and past examination question papers, the performance of the candidates, statistics, and other information[vi].

For some years there has been interest in finding innovative methods of recommending grade boundaries (e.g. French et al, 1992), or improving current practices (e.g. Cresswell, 1997). However, in the past five years modernisations such as e-marking[vii] (for more information see, Price and Petre, 1997; Whetton and Newton, 2002; Raikes and Harding, 2003; Sturman and Kispal, 2003; Leacock and Chodorow, 2003; Sukkarieh et al, 2005; Fowles and Adams, 2005), have renewed interest in innovative methods of recommending grade boundaries, e.g. Thurstone pairs (Pollitt and Elliott, 2003a and b), 'remote Awarding' (Meyer et al, 2006) and rank ordering (Black and Bramley, in press). Additionally, the DfES[viii] and QCA[ix] in association with EdExcel[x] and AQA are funding Kimbell et al (2007) to experiment with a mixture of rank ordering and Thurstone pairs, as well as to investigate the challenges that Awarding Bodies[xi] would face if this approach replaced current Awarding.

The purpose of this paper is to establish which of the above methods successfully avoids the limitations and plays on the strengths of human judgements about the quality of candidates' work. To this end, the available research evidence (about the Awarding methods) is used to evaluate each method against the same criteria. Thus far, no-one has publicly scrutinised the methods in this way. This paper is also intended to contribute to the current public debate about appropriate methods for Awarding. Clearly, this paper is an opinion based on research literature and so colleagues might have different views.

2. Scope of the research

When judgements are made about whether candidates' work is of equivalent quality from one year to the next judges[xii] are expected to judge the performance of the candidates whilst taking into account the demand and difficulty of the current and previous examination question papers[xiii]. We call this 'judging the quality of candidates' work', and this complex process is the focus of my paper.

A great deal of psychological research shows that people, including experts, do not generally reason well using probabilities or statistics (Gigerenzer, 2002). Gigerenzer (2002) explains in detail which ways of presenting statistical information can improve how well people reason with the information. Research evidence showed that some judges misunderstood or misinterpreted some of the statistics provided at Awarding meetings[xiv] (Cresswell, 1997; Murphy et al, 1995). Cresswell (1997) argues that judgements about the quality of candidates' work can be swayed by the statistical information provided at Awarding meetings. It can be inferred from the above that judges' expertise is best used to judge only the quality of candidates' work. In other words perhaps judges (subject experts) should judge only the quality of candidates' work to recommend a grade boundary or a range of marks within which the boundary should be set. Subsequently, any necessary statistical processing or decisions could be undertaken by statisticians in a separate procedure. On the other hand there are some assessment professionals who maintain that once a syllabus has been established only statistical methods are required to set grade boundaries. Thus, the focus of this research is judging the quality of candidates' work.

The main research question is:-

Which Awarding procedure successfully avoids the limitations and plays on the strengths of human judgement? This question is considered within the context of judging that the quality of candidates' work on grade boundaries is of the same quality from one year to the next. The sub questions are:-

1) What are the principles, strengths and limitations of human judgement?

2) What are the circumstances that avoid the limitations and play on the strengths of human judgement?

3) How well do the 'current Awarding', 'remote Awarding', Thurstone pairs and rank ordering procedures avoid the limitations and play on the strengths of human judgement for the purpose of judging the quality of candidates' work?

In this paper each question will be addressed. Although the focus of the research is narrow it will provide an in-depth understanding of some widely used processes.

Other Awarding procedures have been suggested by Greatorex (2003), but they are at the item level rather than the examination level, so they are beyond the scope of this paper. For a summary of why examination level rather than item level approaches for standard setting are preferable for A-levels and GCSEs in England see Black and Bramley (in press). Additionally, it is beyond the scope of this paper to consider changes to A-levels and GCSEs such as pre-testing and associated practices (see Baker et al (2002) and /or Black and Wiliam (2002) for further details).

In this paper I take a psychological view point which explains how people make judgements *in general*. However, a limitation of my paper is that, it some judgements might not conform to these general trends. In some of the Awarding procedures mentioned above these judgements can be identified.

3. Principles, strengths and limitations of human judgements

A current psychological understanding of the principles, strengths and limitations of human judgement is briefly outlined below. It is difficult to do justice to this field in a few words; however, an in-depth review is beyond the scope of this research.

Work by Dr Laming whilst at the University of Cambridge, is included. He synthesised a century of psychological research and drew from his own experiments to develop a theory of human judgement.

Laming (2004, 51) provided evidence that all judgements are a comparison of one thing with another, and these judgements are "little better than ordinal". Additionally, people cannot hold an accurate and stable frame of reference in memory. What is more, humans can only distinguish five categories on a given continuum, if they have no support. Consequently, when people make a series of judgements (about a similar topic), their recent judgements can be influenced by their own earlier judgements and memories can become confused. Unsurprisingly, this can lead to errors of judgement. (Although Laming did not mention it, these errors are sometimes known as 'order effects'). Consequently, judgements are more accurate when they are ordinal or binary comparisons with a reference point or scale rather than absolute judgements using internal standards. From this we can infer that ideally in an Awarding procedure *judgements are an ordinal or binary comparison with a reference point or scale; therefore no absolute judgements are required (criterion 1).*

Judgements can be swayed by group dynamics, which are very powerful. Sometimes people make incorrect judgements to fit in with the crowd, even when it should be clear that the general view is inappropriate. Additionally, the prestige of a person within a group influences the weight that is given to their judgement or views (ibid, 2004). Consequently, when a person of high prestige makes an

inappropriate judgement others might follow suit. Hence, making judgements individually rather than as a group or in a meeting seems to be a situation which enables humans to make better judgements. Ideally then, in an Awarding procedure, *judges make decisions independently of other judges (criterion 2)*.

Human judgements, even expert judgements, can be swayed by extraneous information which should be judged to be superfluous. For example, expert judgement about the quality of art or literature is heavily influenced by the presumed authorship rather than the quality of the writing or painting (ibid, 2004). Arguably, people make better judgements when they are not aware of extraneous information. It follows that if at all possible, in an Awarding procedure *the presence of extraneous information (and its effects) is minimised (criterion 3)*. Clearly, if judges are taking extraneous information into account then they are not making decisions based on an appropriate construct. This should be avoided in an Awarding procedure, and is an issue of construct validity[xv].

Prior experience is often used as a reference point, and past experience makes judgements more repeatable (ibid, 2004). Therefore, one way of making judgements more repeatable is by using judges with appropriate experience. It is desirable that decisions about examination grade boundaries are repeatable so that they are trustworthy. It appears that preferably in an Awarding procedure *judges have appropriate experience (criterion 4)* and *the decisions are reliable (repeatable) (criterion 7).*

Humans make more accurate decisions when they make a series of atomistic judgements rather than one holistic judgement (Dawes and Corrigan, 1974). For example, many studies have found that a linear (statistical) combination of diagnostic signs (identified by clinicians) is more accurate than the clinicians' overall judgement (Laming, 2004). Research also demonstrates that experts are good at knowing what to look for, but they are not good at mentally combining information (Dawes, 1979; Einhorn, 2000; Laming, 2004). If we apply these arguments to the Awarding context it seems that experts know what qualities to heed in candidates' work, and what construct to measure. But, the judges might not be as good at integrating information (e.g. combining statistics with the quality of candidates' work, or mentally aggregating the qualities of candidates' work from different examinations). It seems that if at all possible in an Awarding procedure *judges are not expected to mentally combine performance information, e.g. from different examination question papers in the same session. (Judgements about different examination question papers are statistically combined) (criterion 5).*

In addition to Laming's work (above) this section covers the seminal work of Tversky and Kahneman (1982) who are psychologists working in the field of the heuristics (mental short cuts) and biases in human judgement. Often human judgement is considered to involve simplifying heuristics (Gilovich, Griffin and Kahneman, 2002). These heuristics draw on complex underlying cognitive processes. Generally these heuristics are successful and result in accurate judgements, however, they can lead to inadvertent biases (Gilovich and Griffin, 2002). For example, Tversky and Kahneman (1982) found that sometimes people evaluate the likelihood of an event by the ease with which occurrences can be brought to mind. They called this mental short cut the *availability* heuristic. Often the availability heuristic is useful because more frequent events are usually recalled more easily than infrequent ones. But the availability heuristic can lead to biased judgements, for example *biases due to the retrievability of instances*. One class might be judged larger than another, even though the two classes are of equal size. The bias is caused because the class whose instances are familiar are more easily recalled. Seemingly, these heuristics are generally successful, so it is sensible not to interfere with their use in Awarding unless there is experiential evidence that biases are occurring.

Making judgements about the quality of candidates' work might involve a good deal of reading and understanding of long texts (Sanderson, 2001; Johnson and Greatorex, 2006). Therefore, research about the reading and understanding of long texts will be considered in this paper, to supplement the literature above. There is research evidence that presenting long texts on screen rather than on paper makes reading and understanding them more difficult (O'Hara and Sellen, 1997; Sellen and Harper,

2002). Greatorex (2004) found that reading e-portfolios was different from and more difficult than reading paper portfolios, due to issues like navigation. Overall it can be deduced from the above research that it might be more difficult to read and understand candidates' work presented on screen than if the same candidates' work were presented on paper. Following this argument, ideally in an Awarding procedure *judges are presented with candidates' work in a mode that facilitates rather than hinders the judges' understanding of what candidates have written (criterion 6).*

Figure 1 Principles, strengths and limitations of human judgement: a summary

A All judgements are a comparison of one thing with another, and these judgements are "little better than ordinal".

B People cannot hold an accurate and stable frame of reference in memory.

C People can only distinguish five categories on a given continuum.

D Recent judgements can be influenced by earlier judgements.

E During a series of judgements memories can become confused.

F Judgements can be swayed by group dynamics.

G Judgements can be influenced by extraneous information.

H Prior experience is often used as a reference point, and past experience makes judgements more repeatable.

I Atomistic judgements are generally more accurate than holistic judgements. That is, people are good at knowing what to look for, but they are not good at mentally combining information.

J People tend to use heuristics, which are usually successful.

K When people read an extended text from screen rather than from paper they use different reading strategies and it can be more difficult to understand the text.

Points A to I are mostly from Laming (2004), point J is from psychologists such as Tverskey and Kahneman (1982) and point K is from O'Hara and Sellen (1997) and Sellen and Harper (2002).

The intention of this literature review was to identify the principles, strengths and weaknesses of human judgement. From this understanding of human judgement we drew criteria that an Awarding procedure should ideally meet. The resulting criteria are below.

Figure 2 Criteria that an Awarding procedure should ideally meet

1. All judgements are an ordinal or binary comparison with a reference point or scale; therefore no absolute judgements are required.
2. Judges make decisions independently of other judges.
3. The presence of extraneous information (and its effects) is minimised. (Judges should make decisions about an appropriate construct).
4. Judges have appropriate experience.
5. Judges are not expected to mentally combine information, e.g. from different examination question papers in the same session. Judgements about different examination question papers are statistically combined.
6. Judges are presented with candidates' work in a mode that facilitates rather than hinders the judges' understanding of what candidates have written.
7. The judgements are reliable (repeatable).

In the next section current Awarding and possible alternative methods that have been suggested in the public domain research literature for UK A-levels and GCSEs will be evaluated against these criteria.

4. Awarding procedures

4.1. Current Awarding practice

In the 'Code of Practice', QCA publish the processes and procedures to which GCSE and GCE A-level practices must adhere (QCA, 2006/7). The summary given below is a description of what generally happens for GCE and GCSE examinations, but there are some procedures which are particular to coursework, multiple choice tests and so on, which are all detailed in the Code of Practice.

Normally, for GCEs or GCSEs the Principal Examiner (PE) writes the examination question paper, marks some of the candidates' work and supervises all additional marking. An Awarding committee of Principal Examiners and other senior examiners make recommendations for grade boundaries. The Awarding committee uses statistics, and their judgement about the demand and difficulty of examination question papers, the performance of candidates, as well as other information, to recommend 'judgementally awarded grade boundaries'. The remaining grade boundaries are determined arithmetically. This process is undertaken for each externally assessed unit (which is usually one examination question paper). Subsequently, the grade boundaries for the qualification are calculated.

The aim of the Awarding committee is to recommend the grade boundaries for each externally assessed unit in line with the grade boundaries of the same examination and qualification from previous years. The only exception is in the first year of an examination or qualification when the standard is set for the first time.

Initially, a committee hears the PE's report[xvi] about how well the examination question paper worked and other information. During the Awarding meeting the committee starts by looking at examples of candidates' work at the top of the range in which the grade boundary is expected to be (e.g. grade A) and note the mark at which it is unclear that the candidates' work has the unique characteristics of a grade A. Then the judges start reading candidates' work at the bottom of the range in which the grade boundary is expected to be and record when it is unclear that the candidates' work is worth a grade B. Normally there is a resulting range of marks, which is known as the zone of uncertainty. More unusually a single mark rather than a range is chosen, in which case this mark is recommended as the grade boundary. It should be noted that when judges study the candidates' work they can also refer to archive candidates' work at the appropriate boundary and to grade descriptors[xvii]. Once the zone of uncertainty has been determined the judges use their "*collective professional judgement*" (QCA, 2006/7) by referring to statistics and other information to recommend the appropriate grade boundary. That the QCA Code of Practice refers to the second stage as being 'collective' implies that finding the zone of uncertainty is an individual activity, but Cresswell (1997) explains that this is not always the case. Sometimes the Awarding committee works individually and sometimes they make comments to one another and confer. This combination of judging the quality of candidates' work and technical considerations is repeated for each unit for all judgementally awarded grades. At the end of the Awarding meeting the examiners recommend grade boundaries for each of the units and the whole qualification to the Awarding Body. A full description of the Awarding process can be found in QCA (2006/7). In this review the current Awarding procedure will be called 'current Awarding' to distinguish it from other procedures.

In other literature current Awarding has been viewed as 'limen referencing', 'soft criterion referencing', and 'cohort referencing'. For a full discussion of these issues see for example Greatorex (2003) or Baird et al (2000). It is also very similar to a procedure described by Livingston and Zieky (1982) as the 'up and down method'.

4.2. Remote Awarding

AQA are currently considering some modernisations to current Awarding, in the form of remote Awarding as described by Meyer et al (2006). In remote Awarding, the judges do not meet face to face, rather they view the candidates' work (presumably as scanned images) and the other information usually provided in the meeting on screen at home. When judgements about the quality of candidates' work are made remotely the decisions are recorded and collated. At the same time the judges consider the other information usually provided in a face to face meeting. Subsequently, there is a virtual meeting when the information about each grade boundary is discussed and recommendations for grade boundaries are made. The meeting is conducted using virtual classroom technology. Meyer et al suggest judging fewer examples of candidates' work in remote Awarding than is the case in current Awarding. Meyer et al suggest a number of ways to reduce the number of scripts scrutinised, one suggestion is that each individual judge might scrutinise a maximum of 4 scripts per judgementally awarded grade boundary. The number of scripts scrutinised in current Awarding depends upon the size of the range of marks that is scrutinised.

Some arguments about current Awarding probably apply to remote Awarding. For example, there is a great deal of information to deal with and judges might feel overwhelmed by the task of integrating such a variety of information. On the other hand the face to face dynamics of meetings might be replaced by virtual meeting dynamics. There is already a great deal of research about how virtual meeting dynamics differ from face to face meeting dynamics in a variety of non-examining contexts (e.g. Chidambaram and Jones, 1993; Anson and Munkvold, 2004). Additionally, reading and comprehending long texts is different for different modes (see section 3). Therefore, it is not clear to what extent research about current Awarding can be generalised to remote Awarding. Remote Awarding is still being trialled by AQA and so there is little research in the public domain about it.

4.3. Thurstone pairs

Thurstone pairs is used as a research method in many recent Awarding Body comparability studies. These comparability studies are intended to compare standards:-

- between different Awarding Bodies (e.g. Forster and Gray, 2000; Arlett, 2003; Greatorex et a 2003; Edwards and Adams, 2002, 2003; Guthrie, 2003);
- over time (e.g. Bramley et al, 1998).

The purposes of the study determine the examples of candidates' work that are used (whether they are clean of marks or not) and the judges who participate. Nonetheless, essentially the method entails senior examiners individually judging many pairs of candidates' work and deciding which **candidate's work in each pair is the best**. The details regarding how the method was operationalised varied somewhat with different studies, e.g. instructions to judges varied (Bramley, in press). However, some aspects of the studies are more consistent, for example, in many studies the judges have all attended one meeting in which they are all asked to make their own individual rather than collaborative decisions about candidates' work. Bramley (in press) explains that once the decisions have been collected they are statistically analysed to put all the candidates' marks from both examinations onto one scale and compare the standards of the different examinations.

Pollitt and Elliott (2003 a and b) suggested that Thurstone pairs could be used as a method for maintaining standards, and they provide a detailed discussion of the advantages and disadvantages of the method.

To operationalise the method, in each pair there should be one candidate's work from the examination to be graded and another from an archive examination. Additionally, there should also be a series of pairs

for comparison. Subsequently, when all the decisions have been collated a statistical analysis can be used to equate the archive boundary mark with a point(s) on the mark scale for the live examination. Essentially this is a judgemental equating exercise. Pollitt and Elliott (2003 a and b) also point out that such exercises could be undertaken remotely and that they do not necessarily need to be undertaken with judges in a meeting. Most recently, Kimbell et al (2007) experimented with using Thurstone pairs as a method of recommending grade boundaries for GCSE course work that was judged on screen. Pollitt and Elliott (2003a and b) and later Kimbell et al (2007) advocate Thurstone pairs as a possible method of recommending grade boundaries in the future. A further stage in the work undertaken by Kimbell et al is investigating the challenges that would be faced by Awarding Bodies if they implemented this suggested change.

4.4. Rank ordering

From a practical perspective one of the main limitations of Thurstone pairs is the large number of comparisons needed to undertake a robust statistical analysis. To overcome this limitation Bramley et al (1998) suggested that small samples of candidates' work, e.g. ten examples, should be rank ordered from the best to the worst, and that this should be repeated for a number of packs of examples of candidates' work. (The candidates' work is cleaned of marks). Subsequently, the judges' decisions are collated and submitted to a statistical analysis. Using the rank ordering approach means that many paired comparisons are simulated so fewer actual comparisons are needed and as a result rank ordering is more efficient than Thurstone pairs for the judges. Rank ordering has been successfully used in experiments to judgementally equate cut scores on two tests (Bramley, 2005; Black and Bramley, in press). Therefore, Black and Bramley (in press) argue that rank ordering could potentially be used as a standard maintaining procedure (if the score from one test is known, the judgemental equating can be used to give an equivalent cut score on the other test). They also suggest that for Awarding purposes the candidates' work could be judged remotely. As with Thurstone pairs, Kimbell et al (2007) experimented with using rank ordering as a method of recommending grade boundaries for GCSE course work that was judged on screen. Kimbell et al (2007) advocate rank ordering as a possible method of recommending grade boundaries in the future. A further stage in their work is investigating the challenges that would be faced by Awarding Bodies if they implemented the suggested change.

Advocates of Thurstone pairs and rank ordering have argued that a major advantage of both methods is that the statistical analysis cancels out the internal standards of the judges (if internal standards are used) as the individual scales that the judges use are all amalgamated into one scale (Elliott and Greatorex, 2002; Bramley, in press). Furthermore, it is an intended feature of Thurstone pairs and rank ordering exercises that participating judges do not refer to information like grade descriptors and statistics, when they are judging candidates' performance. This is exemplified in a number of Thurstone pairs exercises (Bramley et al, 1998; Forster and Gray, 2000; Arlett, 2003; Greatorex et al, 2002, 2003; Edwards and Adams, 2002, 2003; Guthrie, 2003), as well as a number of rank ordering studies(e.g. Bramley, 2005; Black and Bramley, in press).

5. Evaluating the Awarding procedures against the criteria

In the following section each of the methods will be evaluated against the common criteria. To reiterate, the focus of this research is judgement about the quality of candidates' work. Additionally, to evaluate the different methods we will be drawing from research (about these methods) which is in the public domain. Initially, the research about current Awarding will be reviewed (since this is the area where most information is available) followed by the research about the other methods.

5.1. Current Awarding

5.1.1. All judgements are an ordinal or binary comparison with a reference point or scale; therefore no absolute judgements are required.

In current Awarding if judges make ordinal or binary comparisons these are most likely to be comparisons of the live examples of candidates' work with the archive. Alternatively, the archive examples of candidates' work can be seen as a reference point. Murphy et al (1995) found little evidence of archive examples of candidates' work being used in Awarding meetings. However, Baird (2000) established that to judge the quality of candidates' work some examiners compare live examples of candidates' work with archive examples. Baird explains that some psychologists call this approach 'similarity judgements'. (Baird drew from psychological work by Rosch and Mervis, 1975). The part of the criterion that no absolute judgements are required is somewhat fulfilled by current Awarding as it involves some similarity judgements or binary judgements. None the less, current Awarding does not make the most of psychological human judgement processes, as some judges are making absolute judgements whilst using their own internal standards. For example, research shows that these internal standards (personal views of what constitutes grade-worthy features of candidates' work) dominate decision making (Murphy et al, 1995). As explained previously Laming theorises that humans cannot maintain fixed internal standards in their heads or use internal standards to make consistent judgements. Indeed, a series of studies concur with this theory in the Awarding context (Good and Cresswell 1988a; Cresswell, 1997; Cresswell, 2000; Baird and Scharaskin, 2002; Scharaskin and Baird, 2000). It seems that the only research with contrary results is that of Baird (2000), which established that some judges have internal standards which are unchanged by reading archive examples of candidates' work. Baird (2000) explains that when judges make judgements they compare the content of the candidates' work with their own mental prototype of what constitutes the features of a script from a particular grade, and that some psychologists would refer to such judgements as 'categorising judgements' (Baird, 2000). Some might argue that judges use grade descriptors as an ordinal scale into which they assign candidates' work for the purposes of recommending grade boundaries. However, Murphy et al (1995) found that the extent to which grade descriptors were used in Awarding meetings was variable. Sadler (1985, 1987, 1989) explains that in order for judges to learn what constitutes a particular standard they must use both written standards as well as examples of work. In short it seems that judges generally do not, and are unlikely to be able to use grade descriptors as an ordinal scale. In summary this method does not fit the criterion regarding the ordinal scale.

As already explained, when humans make successive judgements their later judgements tend to be affected by their earlier judgements as human memories can become confused or amalgamated (Laming, 2004). It is likely that current Awarding is not immune to this phenomenon (see below).

Cresswell's (1997) research led him to theorise that judges draft and redraft their understanding of what features constitute an example of candidate's work at a particular grade as they read and judge the quality of more candidates' work. This seems to concur with Laming's theory. To judge the quality of candidates' work it is likely that judges remember the standards from their experience of previous examinations and teaching, and / or archive examples of candidates' work viewed earlier in the meeting. Arguably, their memory of the archive examples of candidates' work (and standards) they have experienced can get confused and / or amalgamated with memories of other candidates' work (and previously encountered standards). Judges might use the availability heuristic so that recently viewed archive examples of candidates' work or teaching experiences are more prominent in their mental prototypes of the features of candidates' work for a particular grade than those viewed previously. It is possible that these most recent experiences could bias judgements, or it could be that the Awarding committee has a variety of recent experiences and so a variety of skills and knowledge will be valued.

As explained above, the Code of Practice for current Awarding asks judges to make judgements in a certain order, which might lead to order effects due to the order in which the tiers are judged, or the order in which the candidates' work at a particular boundary is judged. There is evidence in the public domain of the former (Good and Cresswell, 1988a), but not of the latter. Apparently, there is no research evidence about a third potential source of order effects in current Awarding, i.e. the order in which boundaries are judged within one examination. However, Laming's theory would predict that there would be an effect.

In summary, there is some evidence that some current Awarding judgements are affected by some (but not all) types of earlier judgements.

5.1.2. Judges make decisions independently of other judges.

The Code of Practice implies that in current Awarding judging the quality of candidates' work should initially be an individual decision. Cresswell (1997) did not find this to be the case; rather he showed that influences common to many decision-making committees, e.g. personality, also influenced Awarding decisions. Clearly, current Awarding does not meet the criterion.

5.1.3. The presence of extraneous information (and its effects) is minimised.

It is of paramount importance that judges are judging an appropriate construct. If judges are taking extraneous information into account then the construct being measured might not be appropriate.

It has already been mentioned that research evidence shows that judgements about the quality of candidates' work in current Awarding are influenced by extraneous information. For example, if judgements are made at the examination level the severity of the judgements is different to when judgements are made at the qualification level (Baird and Scharaschkin, 2002). Additionally, the consistency of candidates' performance shapes judges' decisions (Cresswell, 1997; Scharaschkin and Baird, 2000), which it should not, as the examinations are designed using a principle of compensation[xviii].

Features of candidates' work such as centre number, candidate's name, initials of the marker (and sometimes therefore their position in the hierarchy), candidate's sex, tidiness of the writing and so on might all sway judgements in any of the methods. There does not seem to be any research about whether these particular factors are part of judgements in current Awarding, so we do not know if it measures up to the criterion.

Arguably an improved situation for judging the quality of candidate's work in any method could be achieved by using candidates' work that has been anonymised and cleaned of marks, which might be possible in an e-marking system.

In summary, current Awarding does not always meet this criterion.

5.1.4. Judges make judgements about an appropriate construct.

Cresswell (1997) and later Crisp (2007) both found that in current Awarding judges were mostly paying attention to valid information when they were judging the quality of candidates' work, e.g. Geography judges attended to Geography skills and knowledge. This suggests that an appropriate construct is being measured during the current Awarding procedure. However, they also provide evidence that sometimes judges paid attention to less relevant information. Cresswell's (1997) results indicated that a some judgements about the quality of candidates' work did not relate to the subject content and that the judgements were not explicitly contextualised within the questions or examination question papers. Crisp (2007) showed that when judges paid attention to less relevant information they sometimes tried to

ensure that these factors did not influence their judgements. It seems that generally judges are judging an appropriate construct.

5.1.5. Judges have appropriate experience

Wiliam (1996) argues that the maintenance of standards requires that standard setters (in this case GCSE and GCE judges and accountable officers) must be full participants in a community of practice and they must be trusted by the users of assessment results. From a pragmatic perspective, the Code of Practice, and Awarding Body procedures outline the qualifications and experience that people need to have to be judges. However, there does not seem to be any research indicating who should be included as judges for current Awarding, or what qualifications or experience they should have.

5.1.6. Judges are not be expected to mentally combine information, e.g. from different examination question papers in one session, so judgements will be made about one examination question paper at a time. Judgements about different examination question papers are statistically combined.

As already mentioned there is evidence that current Awarding judgements about the quality of candidates' work are influenced by whether judgements are made at the subject or examination question paper level (Baird and Scharaskin, 2002). Perhaps, this might be because judges expect less from the candidates' performance at the subject level (Baird and Scharaskin, 2002). Alternatively, it could be that the judges know what to look for in the candidate's work to make judgements about quality but that they are not so good at mentally combining the information about different examinations.

In current Awarding judgements are made about each examination question paper at a time. Subsequently, these decisions are statistically combined. In this regard the criterion is met for current Awarding.

Research shows that sometimes in current Awarding the judges experienced difficulties in synthesising the large amounts of information they needed to use to make decisions (Cresswell, 1997; Murphy et al, 1995). In other words a disadvantage of current Awarding is that the judges have to mentally combine a variety of information (statistics, evidence from candidates' work, grade descriptors, PE's reports etc.). Therefore, current Awarding does not meet the criterion, in this regard.

5.1.7. Judges are presented with candidates' work in a mode that facilitates rather than hinders the judges' understanding of what candidates have written.

Candidates' work is viewed on screen (rather than on paper) in a minority of current Awarding meetings. It is likely that presenting candidates' work that includes long texts on paper (rather than on screen) facilitates reading the work (and judging the quality of candidates' work) using strategies that help text comprehension. Additionally, reading or judging the quality of candidates' work on screen might be a qualitatively different process to judging the same work on paper, this might lead to a qualitatively different construct being measured when work is viewed on paper versus on screen. There seems to be no research evidence to say what the situation is for current Awarding so it is not clear whether this criterion is fulfilled.

5.1.8. The judgements are reliable (repeatable).

The reliability of judgements about the quality of candidates' work is shaped by a number of factors (further details are given above). The precision of current Awarding is less than perfect, and this situation has been apparent for some time. For example, see Willmott and Whittall's (1975) work about the predecessor qualifications of GCSE – the General Certificate of Education O Levels and the Certificate of Secondary Education. In some more recent research Good and Cresswell (1988b) replicated some current Awarding meetings for French, History and Physics. Good and Cresswell (1988b, 23 in Cresswell, 2000, 63) concluded that "different groups of grade awarders can reach decisions about final grade boundaries which are sufficiently similar to be acceptable, given the inherent imprecision of the examining process." There is a good deal of evidence to suggest that the results of current Awarding are reliable, therefore the criterion is met by current Awarding.

5.2. Evaluating the alternative suggested Awarding methods

Whilst reviewing the research literature about Thurstone pairs, rank ordering and remote Awarding it became apparent that there was a lack of research literature that could be used to evaluate the methods against the common criteria. A summary of the evidence that is available will be given, followed by an outline of the research which should be undertaken. Once such research is completed it should be possible to establish which method of Awarding best meets the common criteria. Subsequently, when decisions are made about which method to use, practice will be underpinned by research evidence.

5.2.1. All judgements are an ordinal or binary comparison with a reference point or scale; therefore no absolute judgements are required.

Thurstone pairs and rank ordering are often advocated as making good use of human judgement processes as judges are not required to make absolute judgements or use internal standards, e.g. Elliott and Greatorex (2002), Black and Bramley (in press). From this perspective these methods fulfil the part of the criterion that no absolute judgements are required. Additionally, an advantage of Thurstone pairs over the other methods is that it would be possible to design Thurstone pairs exercises that guard against order effects, for instance, by judging candidate's work in a random order. (It is difficult to imagine how the other methods could be adjusted to guard against order effects).

5.2.2. Judges make decisions independently of other judges.

Regarding remote Awarding, some of the group dynamics from face to face meetings might be reduced (Meyer et al, 2006), although different virtual meeting dynamics might be evident instead, however, there appears to be no research on this issue. Meyer et al's outline of remote Awarding suggests that the judgements of the quality of candidates' work will take place remotely and individually. Subsequently, there is a virtual meeting with the aim of recommending the grade boundaries. During the virtual meeting there is some discussion about the collated judgements and other information that is usually used in current Awarding. So it appears that remote Awarding is intended to meet the criterion that judges should make the decisions about candidates' work independently of other judges.

It has been found that the results from a postal rank ordering study were replicated in a repeat study run as a face to face meeting (Black and Bramley, in press). From this we can deduce that there are few face to face meeting dynamics in rank ordering studies that influence judgements, and / or that rank ordering studies can be successfully conducted by post. During rank ordering and Thurstone pairs face to face meetings, the judges were asked to make their decisions individually, although this might be difficult to enforce. It is expected that undertaking rank ordering or Thurstone pairs by post will increase

the likelihood that decisions are made independently. So the rank ordering and Thurstone pairs methods are intended to meet the criterion that judges should judge the quality of candidate's work independently of other judges.

5.2.3. The presence of extraneous information (and its effects) is minimised.

In current Awarding the judges can see the marks credited to the candidate's work, and this is a potential source of extraneous information. For example, Scharaskin and Baird (2000) found that the consistency of performance within scripts affects judgements of gradeworthiness. When marks are not available in rank ordering studies they cannot influence decisions (Black and Bramley, in press). Likewise when the marks are not available in Thurstone pairs, the marks cannot influence decisions. In this regard Thurstone pairs and rank ordering meet the criterion.

It is inherent in the rank ordering method that the judgements are not contaminated by statistics (Black and Bramley, in press). This argument also applies to Thurstone pairs, but not to current Awarding or remote Awarding. In this way rank ordering and Thurstone pairs best meet the criterion.

5.2.4. Judges make judgements about an appropriate construct.

In some Thurstone pairs exercises the judges were asked what they had been attending to (Edward and Adams, 2002, 2003). Some examples of the responses are the complexity/range of skills, breadth and depth of knowledge and the use of case study material (Edward and Adams, 2002). Whilst these studies list only some of the features of the candidates' work that the judges reported to be using, the judges do seem to be attending to a relevant construct. Therefore, Thurstone pairs seems to meet the criterion.

5.2.5. Judges have appropriate experience

In the Thurstone pairs procedure there are sometimes some independent judges[xix] as well as people who are qualified to be judges in current Awarding. Forster and Gray (2000) found that the judgements of independent judges were not statistically different to the judgements of the other judges. The independent judges made fewer judgements than the other judges, perhaps this is due to lack of familiarity of making judgements about candidates' work. It seems that for Thurstone pairs some independent judges do have appropriate experience to be involved in such exercises.

5.2.6. Judges are not be expected to mentally combine information, e.g. from different examination question papers in one session, so judgements will be made about one examination question paper at a time. Judgements about different examination question papers are statistically combined.

Research shows that sometimes in current Awarding the judges experienced difficulties in synthesising the large amounts of information they needed to use to make a decision (Cresswell, 1997; Murphy et al, 1995). The advantage of rank ordering and Thurstone pairs in comparison with current and remote Awarding is that the judges do not have to mentally combine a variety of information (statistics, evidence from candidates' work, grade descriptors, PE's reports etc.). Therefore, rank ordering and Thurstone pairs meet the criterion but the other two methods do not. All of the methods can be organised so that judgements are made about one examination paper at a time, and that subsequently statistics are used to combine the judgements.

5.2.7. The judgements are reliable (repeatable).

The reliability of the underlying scale in Thurstone pairs studies has been demonstrated by Jones and Meadows (2004). They repeated a study using Thurstone paired comparisons in GCSE Religious Studies using a second set of judges but the same candidates' work. The correlations between the judgements in the two different studies were high.

The trait measured in a rank ordering study was linearly related to the marks (Bramley, 2005). When a postal rank ordering exercise was repeated in a further study as a meeting based exercise the results of the first exercise were replicated, illustrating that the method is reliable. The correlations between the judgements in the two different conditions were 0.93 for one experiment and 0.91 for the other experiment (Black and Bramley, in press). In the same research the rank ordering generally correlated with the mark order.

Kimbell et al (2007) used a combination of rank ordering and Thurstone pairs. They claim that in this combination “The standard error attaching to the placement of individuals with the rank order is significantly lower than would be the case in conventional portfolio assessment” (Kimbell et al, 2007, 6).

There is a good deal of evidence to suggest that the results of rank ordering and Thurstone pairs are reliable (repeatable), the criterion is met by these methods and there is little to choose between them in this regard.

5.3. Areas for future research

It can be seen from the above that there is already a relevant and sound body of knowledge developing about each of the alternative Awarding methods. However, there is also a good deal of research that needs to be done before research evidence can be used to evaluate all the methods against the common criteria. If we are to develop the body of knowledge in this area and facilitate evidence based practice a number of research questions remain. These questions are outlined below:

- What types of cognitive judgements (e.g. similarity judgements or comparisons) are used by judges in remote Awarding, Thurstone pairs and rank ordering? Bramley (in press) has alluded to the need for such research in rank ordering.
- Does each judge's memory of examples of individual candidate's work get confused or amalgamated with their memories of other individual's work during Thurstone pairs, remote Awarding or rank ordering?
- Do the judgements made in Thurstone pairs, remote Awarding and / or rank ordering suffer from order effects?
- Does extraneous information (e.g. tidiness, candidate's name, candidate's sex, spread of marks or consistency of performance) influence decisions in remote Awarding, Thurstone pairs and /or rank ordering? Bramley (in press) has alluded to the need for research on similar topics for rank ordering.
- What information (both relevant and irrelevant) is heeded when judges are involved in remote Awarding and rank ordering?
- What construct is being measured in remote Awarding and rank ordering? Bramley (in press) has argued that there is a need for research on similar topics for rank ordering.
- What experience do judges need to have to be involved in recommending grade boundaries using any of these methods?
- How do judges mentally deal with the information that they need to synthesise in remote Awarding? (For example, do judges use mental short cuts or biases?)
- Should candidates' work be presented on paper or on screen, for any of the methods?
- How reliable are the judgements that are made in remote Awarding?

6. Conclusions

It is important to emphasise that the current Awarding practice meets the regulatory criteria set by the Qualifications and Curriculum Authority. In the context of commitment to continuing to explore new approaches researchers, QCA, DfES, and the Awarding Bodies are experimenting with other innovative methods of Awarding (Pollitt and Elliott, 2003a and b; Black and Bramley, in press; Kimbell et al, 2007).

Thus far some authors (including myself) have argued that the advantage of Thurstone pairs or rank ordering is that judges are not required to make absolute judgements, whereas current Awarding rests on absolute judgements (e.g. Elliott and Greatorex 2002; Pollitt and Elliott 2003a and b; Black and Bramley, in press). Arguably, our assertion about current Awarding is overstated. After all, Baird (2000) illustrated that in current Awarding judges use both comparisons and absolute judgements. What is more, research questions remain regarding whether judges use internal standards in rank ordering or Thurstone pairs or remote Awarding. Once this research is completed discussions about whether Thurstone pairs or rank ordering are better than current Awarding can include points about the types of judgements that the judges use.

Many of the advantages of particular Awarding methods are still a matter of opinion rather than research evidence. This highlights the need for further research about issues like the construct validity and judges' cognition for each and every method. To date much of the research has been about these issues in **only** some of the methods, e.g. the features attended to by judges in current Awarding and Thurstone pairs but not rank ordering or remote Awarding. Until recently much of the research associated with these methods was about other issues such as the comparability of standards per se rather than about the methods themselves. However, there is an ongoing programme of research at Cambridge Assessment about methods of recommending grade boundaries (for completed research see Pollitt and Elliott, 2003a and 2003b, Bramley 2005, Black and Bramley in press). Arguably, the issue of what examiners are heeding (the construct being measured) is the most important question to be answered, something to which Bramley (in press) has alluded. It would be useful to target future research at investigating what judges are heeding during rank ordering (Bramley, in press). What judges pay attention to in Thurstone pairs is only a minor strand of some comparability studies. But what judges pay attention to in current Awarding is thoroughly researched, comparatively speaking. Consequently, we are currently undertaking an experiment to investigate for rank ordering, Thurstone pairs and current Awarding (1) what information judges attend to (in other words what construct is being measured) and (2) what cognitive approaches are being used. It is anticipated that our completed and ongoing research will enable us to better evaluate the methods and contribute to the public debate.

Apparently, current Awarding and Thurstone pairs have the advantage that there is research evidence that they seem to be measuring an appropriate construct. Additionally, Thurstone pairs and rank ordering are better by design as they have the advantage that the judgements about the quality of candidates' work should not be swayed by statistics, group dynamics, other judge's opinions or mentally combining information from a variety of sources. (Note, however, that in all the methods judges combine information from different examination questions). Current Awarding has the disadvantage that it can suffer from order effects (Good and Cresswell, 1988a). Thurstone pairs has the advantage over the other methods that it could be organised to guard against order effects.

In summary, it seems that Thurstone pairs has more advantages than any other method, but that many of its advantages are shared with rank ordering. Whilst this paper has come from a particular psychological view the conclusions are similar to those of previous authors like Pollitt and Elliott (2003a and b) and Kimbell et al (2007).

Clearly, in deciding upon a method of Awarding there are not only psychological advantages and disadvantages to consider. Black and Bramley (in press) have discussed some of the practical advantages and disadvantages as well as some of the issues discussed here.

7. References

Anson, R. and Chipmunk, B. E. (2004) Beyond Face to Face: A Field Study of Electronic meetings in Different Time and Place Modes, *Journal of Organizational Computing and Electronic Commerce, 14*, (2), 127-152.

Arlett, S. J. (2003) *A Comparability study in VCE Health and Social Care, Units 3, 4 and 6: a review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2002 examination and organised by AQA on behalf of the Joint Council for General Qualifications.*

Baird, J. (2000) Are examination standards all in the head? Experiments with examiners' judgements of standards in A level examinations. *Research in Education*, *64,* 91-100

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

Baird, J. and 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

Baker, E. et al (2002) *Maintaining GCE A level Standards*, London, QCA. www.internationalpanel.org.uk

Black, B. and Bramley, T. (in press) An investigation and cross-validation of 2004 and 2005 standard setting in GCE A-level Psychology using a rank-ordering method. *Research Papers in Education.*

Black, P. and Wiliam, D. (2002) *Standards in Public Examinations*, London, King's College Department of Education and Professional Studies.

Bramley, T. (2005), A Rank-Ordering Method for Equating Tests by Expert Judgement, *Journal of Applied Measurement, 6,* 2, 202-223.

Bramley, T. (in press) Paired Comparison Methods, *Techniques for monitoring the comparability of examination standards*, A QCA book.

Bramley, T., Bell, J. F. and Pollitt, A. (1998) Assessing changes in standards over time using Thurstone paired comparisons. *Education Research and Perspectives, 25,* 2, 1-23

Chidambaram, L. and Jones, B. (1993) Impact of Communication medium and computer support on group perceptions and performance: A comparison of face to face and dispersed meetings, *MIS Quarterly, 17,* 4, 465-491.

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

Cresswell, M. (2000) Defining, Setting and Maintaining Standards in Curriculum-Embedded Examinations: Judgemental and Statistical Approaches. In H. Goldstein, and T. Lewis, Editors *Assessment: Problems, developments and statistical issues*. (pp. 57 to 84). Chichester: John Wiley and Sons.

Crisp, V. (2007) *Do assessors pay attention to appropriate features of student work when making assessment judgements?* A paper to be presented at the International Association of Educational Assessment annual conference. Baku, September 16~21.

Dawes, R. and Corrigan, B. (1974) Linear models in decision making. *Psychological Bulletin, 81,* 95-101.

Dawes, R. (1979) The Robust Beauty of Improper Linear Models in Decision Making, *American Psychologist, 34,* 7, 571-582.

Edwards, E. and Adams, R. (2002) *A Comparability Study in GCE Advanced Level Geography Including the Scottish Advanced 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 WJEC on behalf of the Joint Council for General Qualifications.*

Edwards, E. and Adams, R. (2003) *A Comparability Study in GCE Advanced Level Geography Including the Scottish Advanced Higher Grade Examinations. A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2002 Examination and organised by WJEC on behalf of the Joint Council for General Qualifications.*

Einhorn, H, J. (2000) Expert judgement: Some necessary conditions and an example, In T. Connelly, H. R. Arkes, and K. R. Hammond (Eds), *Judgement and decision making: an interdisciplinary reader* (2nd edition., pp. 324-335) Cambridge: Cambridge University Press.

Elliott, G. and Greatorex, J. (2002) A fair comparison? The evolution of methods of comparability in national assessment, *Educational Studies, 28,* 3, 253-264

Forster, M. and Gray, E. (2000) *Impact of Independent Judges in comparability studies conducted by Awarding Bodies,* A paper presented at the British Educational Research Association Annual Conference, Cardiff University, September.

Fowles, D. and Adams, C. (2005) *How does assessment differ when e-marking replaces paper-based marking?* Paper presented at the IAEA Conference Abuja, Nigeria. Retrieved from www.iaea.info/abstract_files/paper_051218101528.doc on 5th February 2006.

French, S. Allatt, P., Slater, J., Vassiloglou, M. and Willmott, A. (1992) Implementation of a decision analytic aid to support examiners' judgements in aggregating pairs of components. *Journal of Mathematical and Statistical Psychology, 45,* 75-91.

Gigerenzer, G (2002) *Reckoning with Risk. Learning to Live with Uncertainty.* London: The Penguin Press.

Gilovich, T., & Griffin, D. (2002) Introduction – heuristics and biases: then and now, In T. Gilovich, D. Griffin, and D. Kahnemann (Eds) *Heuristic and biases: the psychology of intuitive judgement*. (pp. 1 to 18) Cambridge : Cambridge University Press.

Gilovich, T. Griffin, D., and Kahnemann, D. (2002) (Eds) *Heuristic and biases: the psychology of intuitive judgement*. Cambridge : Cambridge University Press.

Good, F. J. and Cresswell M. J. (1988a) Grade Awarding Judgements in differentiated examinations. *British Educational Research Journal, 14,* 3, 263-281

Good, F. J. and Cresswell M. J. (1988b) Grading the GCSE. London: Secondary schools Examination Council. In Cresswell, M. (2000) Defining, Setting and Maintaining Standards in Curriculum-Embedded

Examinations: Judgemental and Statistical Approaches. In H. Goldstein, and T. Lewis, (Eds) *Assessment: Problems, developments and statistical issues*. (pp. 57 to 84). Chichester: John Wiley and Sons.

Greatorex, J. (2003) *What happened to limen referencing? An exploration of how the Awarding of public examinations has been and might be conceptualised*, A paper presented at the British Educational Research Association Conference, 10 - 13 September 2003 at Heriot-Watt University, Edinburgh

Greatorex, J. (2004, December) *Moderated e-portfolio project evaluation*. (Evaluation and Validation Unit, University of Cambridge Local Examinations Syndicate). Retrieved May 2007, from the Cambridge Assessment website: http://www.cambridgeassessment.org.uk/research/confproceedingsetc/moderatedEportfolioProjectEvaluation

Greatorex, J., Elliott, G. and Bell, J. F. (2003), *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, University of Cambridge Local Examinations Syndicate for OCR on behalf of the Joint Council for General Qualifications.*

Greatorex, J. , Hamnett, L. and Bell, J. F. (2003) *A comparability study in GCE Chemistry Including the Scottish Advanced Higher Grade. A study based on the Summer 2002 examination and organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for OCR on behalf of the Joint Council for General Qualifications.*

Guthrie, K. (2003) *A Comparability Study in GCE Business Studies and VCE Business, A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2002 Examination and organised by the EdExcel on behalf of the Joint Council for General Qualifications.*

Johnson, M. and Greatorex, J. (2006) Judging learners' work on screen: Issues of validity. *Research Matters: A Cambridge Assessment Publication, 2,* 14-17.

Kimbell, R., Wheeler, A., Miller, S. and Pollitt, A. (2007), *E-scape portfolio assessment phase 2 report*. Department of Design, Goldsmiths, University of London.

Laming, D. (2004) *Human judgement The Eye of the Beholder*, Cambridge: Cambridge University Press.

Leacock, C. and Chodorow, M. (2003) C-rater: Automated Scoring of Short-Answer Questions, *Computers and Humanities, 37,* 4, 389-405.

Livingston, S. A. Zieky, M. J. (1982) *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton: Educational Testing Service.

Murphy, R., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. and Gower, R. (1995) *The dynamics of GCSE Awarding*. Report of a project conducted for the School Curriculum and Assessment Authority, School of Education, University of Nottingham.

Meyer, L., Baird, J., Stringer, N. O'Sullivan, L. and Adams, C. (2006) *Awarding in the 21st century,* A paper presented at International Association for Educational Assessment, May, Singapore.

O'Hara, K. and Sellen, A. (1997). A comparison of reading paper and online documents. *Proceedings of the Conference on human factors in computing systems (CHI '97)*, (pp. 335–342). New York: Association for Computing Machinery.

Pollitt, A., Ahmed, A. and Crisp, V. (in press) *Techniques of exploring the demands of syllabuses and question papers*. *Techniques for monitoring the comparability of examination standards*, A QCA book.

Pollitt, A. and Elliott, G. (2003a) *Monitoring and Investigating comparability: a proper role for human judgement.* Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', UCLES 4th April 2003.

Pollitt, A. and Elliott, G. (2003b) *Finding a proper role for human judgement in the examination system.* Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', UCLES 4th April 2003.

Price, B. and Petre, M. (1997) *Teaching Programming through Paperless Assignments: an empirical evaluation of instructor feedback*, Centre for Informatics Education Research, Computing Department, Open University, UK. Retrieved April 20, 2005, from http://Mcs.open.ac.uk/computing papers/mzx/teaching.doc

Qualifications and Curriculum Authority (2006/7) *GCSE, GCE, VCE, GNVQ and AEA Code of Practice.* London: Qualifications and Curriculum Authority.

Raikes, N. and Harding, R. (2003) The Horseless Carriage Stage: replacing conventional measures. *Assessment in Education, Principles, Policies and Practices, 10,* 3, 267-277.

Rosch, E., and Mervis, C. B. (1975) Family resemblance studies in the internal structure of categories, *Cognitive Psychology, 7,* 573-605.

Sadler, D. R. (1985) The origins and functions of evaluative criteria. *Educational Theory, 35,* 285-297

Sadler, D. R. (1987) Specifying and Promulgating achievement standards. *Oxford Review Education, 13,* 191-209.

Sadler, D. R. (1989) Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119-144.

Sanderson, P. J. (2001) Language and Differentiation in Examining at A level, PhD thesis, School of Psychology, University of Leeds.

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

Sellen, A. and Harper, R. (2002). *The Myth of the Paperless Office.* Cambridge, MA: MIT Press.

Sturman, L. and Kispal, A. (2003, October) *To e or not to e? A comparison of electronic marking and paper-based marking.* Paper presented at the 29th International Association for Educational Assessment Annual Conference, Manchester, UK.

Sukkarieh, J. Z., Pulman, S. G. and Raikes, N. (2005), Automatic marking of short free text responses, *Research Matters A Cambridge Assessment Publication, 1,* 19-22.

Tversky, A. and Kahneman, D. (1982) Judgement under uncertainty: Heuristics and biases, In D. Kahneman, P. Slovic, and A. Tversky. (Eds) *Judgement under uncertainty: Heuristics and biases*. (pp.3-22) Cambridge: Cambridge University Press. Cambridge.

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

Willmott, A. S. and Nuttall, D. L. (1975) *The reliability of examinations at 16+.* Schools Council Research Studies. Schools Council Publications. London: MacMillan Education Ltd.

Whetton, C. and Newton, P. (2002, September). *An evaluation of on-line marking*. Paper presented at the 28th International Association for Educational Assessment Conference, Hong Kong SAR, China.

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

[ii] GCSEs are normally taken by 16 year olds in England and Wales. They are often a prerequisite for GCE level study. The first GCSEs were taken in 1988. They replaced O-levels and Certificates of Secondary Education (CSEs).

[iii] A grade boundary is the lowest mark that a candidate must achieve to get a particular grade.

[iv] The *difficulty* of an examination question or an examination is a measure, often expressed as a facility value. It is usually calculated as a function of the number or percentage of marks a group of examinees obtained. Difficulty is a concept that is applied to groups of students, and does not apply to individual examinees. The difficulty of a question can vary for different groups of candidates (Pollitt et al, in press).

[v] *Demands* are the cognitive demands placed on candidates when they answer an examination question or similar. Demands are a qualitative feature of an examination question (or similar) which *cannot* be measured empirically from students' performance. The demands have to be judged by appropriate experts (Pollitt et al, in press).

[vi] The Code of Practice outlines the numerous sources of information that the judges should use to make decisions. These include, examination question papers and mark schemes, reports from the Principal Examiners about how well the examination question paper functioned, samples of current and archive examples of candidates' work, any published grade descriptors, mark distributions from the current and previous examinations, details of changes in entry patterns, centres' estimated grades for candidates.

[vii] Sometimes, for the purposes of marking, candidates' work on paper can be scanned and presented on screen for examiners to mark.

[viii] The Department for Education and Skills. The predecessor of the Department for Children, Schools and Families, and the Department of Innovation, Universities and Skills.

[ix] The Qualifications and Curriculum Authority is a public body which regulates qualifications offered by Awarding Bodies. QCA are sponsored by the Department for Children, Schools and Families.

[x] EdExcel and AQA are two of the three Awarding Bodies that offer GCSEs and A-levels in England, (OCR is the third).

[xi] Awarding Body is the term used for test agencies that offer GCE A-levels and GCSEs in England.

[xii] The judges, who are responsible for writing the questions in the examination question paper and for leading the marking, are also experienced teachers and examiners.

[xiii] Once the standard for a GCSE or GCE A-level has been set, the intention is to maintain that standard from one year to the next. The examination question papers from each year are different and they are not generally generated from an item bank.

[xiv] The Awarding meeting is when a committee of senior examiners meet to recommend grade boundaries to the Awarding Body.

[xv] Whether a scale measures the construct that it purports to measure.

[xvi] In a modern assessment system, when item level data are available at the Awarding meeting, it might be possible to use statistical information about items in the PE's report.

[xvii] Note that these are grade descriptors which describe typical performance at a grade and therefore allow for different routes to the same grade. They are not grade criteria which have to be met for a candidate to reach a particular grade. As these are grade descriptors and not grade criteria the research about grade criteria for the GCSE is not relevant. Regarding using grade criteria as a method of assessment and grading, Cresswell (2000) is of the opinion that this approach does not work.

[xviii] The principle of compensation means that candidates can gain marks for their strengths without losing marks for their weaknesses.

[xix] A judge who does not have an allegiance to any particular Awarding Body.