What affects the reliability of ‘vocational’ examiners’ marking?

Matthew James Powell-Howard

MA in Education (by research)

The University of York

Department of Educational Studies

September 2009

ABSTRACT

The results of examinations need to be accepted as just, appropriate and reflect the ability of the candidates sitting an examination, and 'examiners' have a key part to play in this process. Examiner's marking to a common standard and a common interpretation of mark schemes is important so as to not disadvantage or favour clusters of students. In addition to disadvantaging or benefitting those sitting an examination, aberrant marking can also affect the integrity of an award and / or qualification by inflating or deflating pass rates.

The purpose of this study was to identify what affects the reliability of vocational examiners marking; with 'vocational' being interpreted as subject experts rather than educationalists. Although there has been extensive research into what affects marking e.g. increased monitoring, clearly structured mark schemes, little research has been undertaken as to what awarding bodies using 'vocational' subject experts have found to be most effective in improving the reliability of the examiners they use.

The assumption of this study was that vocational examiners would need managing differently in terms of selection, training, support and moderation so as to affect the reliability of their marking. What became evident was that the examiners used by the participants in this

study needed no more or less management than *any* examiner, working for *any* examination board or professional organisation offering public examinations, regardless of their background i.e. not being experienced educationalists. The study identified that whenever examiners are being used and from whatever field e.g. experienced examiners, teachers, lecturers etc, they are all potentially fallible and that they need support, guidance and monitoring to be able to fulfil the task of examining reliably and effectively.

CONTENTS

ACKNOWLEDGEMENTS

I would like to thank:

The participants in the study who gave their time willingly, and for their frankness and honesty when giving of their opinions.

My academic Supervisor, Dr Vanita Sundaram, for her time and support.

Frank Stoner for his good advice, clear guidance and for setting me on the road to actually recognising what a research dissertation actually is.

My employer (you know who you are!) for sponsoring me in my studies.

My children Theo, Eden and Darcy, thank you for dropping in to see how I was getting on from time to time, you are all fantastic.

My wife Lex for her continued patience, encouragement, calming influence and advice – you don’t know how good you are and how much I appreciate the help that you have given me.

And finally to my dog Ozzie, you can have that walk now…….!

Introduction

Introduction

> "The aim of all those involved in producing, delivering, assessing, awarding, certificating and regulating accredited qualifications is to make sure that all candidates receive the results their performance merits when judged against the relevant specification content and assessment criteria" (GCSE, GCE and AEA Code of Practice, Qualifications and Curriculum Authority, 2008, p. 4).

My area of enquiry is related to what affects the reliability of vocational examiners involved in the marking of public examinations. By 'vocational' I mean subject experts rather than educationalists. I am interested in this area, as during my time as both an examiner and as a Standards Officer for an awarding body specialising in vocational awards, I was aware that there were variances in how examiners interpreted and applied mark schemes and therefore how they marked candidates work.

Examiners' marking to a common standard and a common interpretation of mark schemes is important so as to not disadvantage or favour clusters of students. In addition to disadvantaging or benefitting those sitting an examination, aberrant marking can also affect the integrity of an award and / or qualification by inflating or deflating pass rates.

Based on personal experience, I am also aware of the lack of training and monitoring examiners undergo. I was considered to be a suitable examiner based purely on my qualifications, relevant vocational experience and Chartered status within my professional field of expertise. My experience (or lack thereof) in relation to marking and educational practices was not considered when appointing me to the role of an examiner nor was any training offered to address this gap in my knowledge and understanding of educational assessment.

It is common practice in the larger examination boards e.g. Edexcel, to use undergraduates to mark core subjects such as mathematics, geography etc however for specialist subjects there is, and should be, a need for subject experts to be used so that they can interpret candidates’ scripts and contribute to what goes onto mark schemes. Individuals who are not necessarily subject experts may mark candidates’ work where they are not required to use a high level of subject expertise to interpret the mark scheme as stated in the Ofqual Code of Practice.

Although there has been extensive research into what affects examiners’ marking performance (generally in relation to those involved in education as a fulltime occupation) e.g. fatigue, the cognitive process of marking and clearly structured mark schemes, little research has been undertaken as to what awarding bodies / professional bodies offering vocational awards and using subject experts have found to be most effective in improving the reliability of the examiners they use. There may also be differing work practices adopted by awarding bodies / professional bodies to ensure that they positively affect the reliability of the examiners that they use.

The document *GCSE, GCE and AEA Code of Practice*, was published by the Qualifications and Curriculum Authority (QCA) in April 2008. The regulatory function of QCA is now the responsibility of the Office of the Qualifications and Examinations Regulator (Ofqual). It is the purpose of Ofqual to regulate qualifications and monitor national curriculum assessments in England, Wales and Northern Ireland. Although the Code is for the awarding bodies that deliver general qualifications (as the title suggests GCSE's, GCE's and the Advanced Extension Award) Ofqual have the expectation that the other awarding bodies that they accredit follow similar processes as set out in the Code, both in terms of setting examinations and the marking of them. Within the examinations sector, accreditation by Ofqual is seen as a badge of quality assurance and by following the guidance given in the Code, best practice can be demonstrable.

The regulators have produced the Code of Practice to meet the public's expectations for high-quality qualifications that are fit for their purpose, command public confidence and are fair and accurate. It sets out the principles of regulation and the criteria for accrediting awarding bodies and qualifications. By referring to the Code I can benchmark the sample used in this study to ascertain if they are exceeding, complying with or performing below what would be expected by the Regulator in terms of monitoring and positively affecting the reliance of the examiners that they use for the marking of their associated awards / qualifications. I will also be examining the research that has been undertaken with regard to what affects the reliability of examiners, some of which, which will have been used to inform the Code.

As a result of my being an examiner, my qualifications and experience of practicing health, safety and environmental management, I was appointed as a Standards Officer for a leading examination board in the field of health, safety and environmental management. The role required me to be responsible for the production and management of a number of examinations at both Level 3 and Level 6 which involved working with a principal examiner(s) in the setting of examination papers and the recruitment and monitoring of examiners.

Whilst engaged in the role of a Standards Officer, it became apparent to me that some examiners were far less reliable than others in terms of the accuracy of their marking and their administration both of which, if not done properly, can have a detrimental or indeed a positive effect e.g. taking an obvious referral to above the pass standard, on a candidates examination results. I also became acutely aware that some of the examiners I was using were not reliable in terms of how they approached marking i.e. their attitude towards candidates and the seriousness with which they undertook the marking process. An example of this was when I routinely witnessed examiners who took it as a personal affront when poor responses were provided by candidates and which provoked, either consciously or subconsciously, into them marking much more harshly than was warranted.

Some examiners found it very difficult to judge when and where to give marks as they could not conceptualise what was required at a specific academic level e.g. levels 3 and 6, and others paid scant regard to command words (those based on the level descriptors given in Bloom’s Taxonomy e.g. list, outline, describe, explain) when looking at the depth and breadth of a response. A common mistake made by examiners was to award marks not contained on a standardised / agreed mark scheme if they felt that a mark was warranted – this in effect meant that examiners all marking the same exam were marking to a different standard.

As previously alluded to poor administration *routinely* causes problems with common errors occurring such as:

- Incorrect adding up of ‘ticks’;
- Transposing the mark given for a question to a mark sheet incorrectly e.g. awarding 5 marks instead of 6; and
- Illegible handwriting.

Very basic in terms of errors but when examiners are marking large numbers of scripts errors do occur which potentially can disadvantage / advantage candidates not to mention cause work for awarding bodies who are already working to tight schedules and deadlines. However if these administration errors are not identified and rectified, the robustness of the examination process can be called into question.

I also found it very difficult to attract and maintain examiners with suitable qualifications and experience of marking as the financial rewards were and still are minimal. The main reason I found examiners were motivated to mark were for a number of reasons. One reason I ascertained, through discussion with them, was so that they had access to mark schemes as a number of them worked for course providers accredited to deliver my organisation’s awards and so they found it advantageous to gain a working knowledge of mark schemes as questions entered a ‘question bank’ and were repeated periodically.

Other motivations appeared to be that individuals wished to be associated with an examination board, for either continued professional development (C.P.D.) reasons i.e. refreshing / gaining knowledge, a requirement of the health and safety professions' principal governing body IOSH (Institution of Occupational Safety and Health), or because they wished to be associated with a respected examination board for curriculum vitae reasons. Another reason which was given repeatedly was that they examined for more philanthropic reasons and that they wanted to contribute to the development of the safety, health and environmental professions and to help maintain standards.

Inaccuracy of marking can be perceived as being problematic for a number of reasons, these being:

- Passes being awarded to candidates who have *not* made the pass standard;
- Passes not being awarded to candidates who *have* made the pass standard;
- Distinctions between candidates i.e. pass, credit, distinction, should be awarded where achieved and in line with the mark schemes set;
- Candidates are able to pay for an enquiry about results if they feel that they have been marked unfairly. Although there will always be instances where scripts have been marked fractionally harshly

around the pass standard e.g. 44% with a nominal pass mark of 45%, in my experience some candidates have had their scripts remarked with variances on occasion exceeding 20% which is wholly inappropriate and of huge concern. Although thankfully this is the exception it is not uncommon for scripts, upon remark, to move both up and down by greater then 5% of the original mark awarded. Successful enquiries about results do bring into question the reliance an examination board / professional body can have over its examiners which in turn can have a detrimental effect in terms of public relations;

- Most, but admittedly not all, candidates for examination undertake a great amount of personal study in preparing for an examination and also at significant financial cost (course providers offering taught courses will charge between £1500 and £8000 for the Level 3 and Level 6 awards offered by the examination boards / professional bodies discussed in this study). It is absolutely appropriate therefore that their work is marked fairly and is open to rigorous scrutiny;
- Poor performance in examinations can have a detrimental effect on an examination board's / professional bodies' growth in terms of appealing to new candidates and may contribute towards students choosing to select other methods of gaining recognised qualifications e.g. National Vocational Qualifications which may

appear to be, rightly or wrongly, more attractive based on pass rates and the avoidance of examinations;

- A course provider's reputation is built on its pass rates on qualifications, it therefore follows that a pass rates should be representative of its true successes / referrals.

As a result of my work as a Standards Officer involved in the monitoring and management of examiners it seemed logical to follow a similar topic of research study although I was not yet clear on how I might approach it. After discussion with colleagues and my research supervisor, I decided to undertake a small-scale qualitative study of examination bodies / professional organisations, who offer qualifications in the field of health, safety and environmental management and use vocational examiners.

In order to research the question 'What affects the reliability of vocational examiners?' Firstly I will review the literature and research available that is specifically focussed on examiners and what affects their performance both in terms of personal characteristics and controls e.g. age, experience level of supervision.

I then aim to discover how three of the most recognised and respected examination boards / professional bodies in the field of health, safety and environmental management manage the use of examiners so as to aid their reliability.

I no longer work in the role of a Standards Officer having moved to another department within the same organisation and although I have a thorough understanding of how an examination board works in terms of its obligations and quality procedures I am aware that things may have changed in how it manages its examiners. I have no knowledge of how the other two organisations selected as part of this study affect the reliability of the examiners they use and in order to understand how all three bodies operate I aim to gather data from those people who have overall responsibility for the performance and management of examinations e.g. a Standards Manager, Director of Awards and a Director of Membership Services. The interviewees selected will be able to give me an understanding of the procedural processes they undertake to influence and check the reliability of examiners marking and also any personal perceptions they have as to what affects a person's ability to mark appropriately. To enable me to gain a balanced perspective and understanding I intend to use semi structured interviews, thereby using qualitative methodology. It is then my intention to transcribe and analyse this data completing a thematic analysis (Denscombe, 1998). I will then present my findings, discussion and conclusions.

There will obviously be ethical issues, these being the potential harm that can be caused to an examination board / professional body should information on how they manage their examiners be exposed to candidates, particularly if the organisation is not seen to be following best practice. As I am a direct employee of one of the organisations (AB2) I would need to keep my boundaries as a researcher, and be very open about my employment to the other awarding body and professional organisation. I must also respect the confidentiality and commercial sensitivity of any documentation, or other information, that I may be privy to.

Confidentiality for the participants was a consideration as was acquiring informed consent, ensuring that all the interviewees understood how the data was to be used and that they would be informed of any use other than that of the dissertation e.g. publication, was confirmed prior to the interviews taking place..

Literature Review

<u>Literature Review</u>

> For ten years I never left my books,
> I went up and won unmerited praise.
> My high place I do not much prize;
> The joy of my parents will first make me proud.
> Fellow students, six or seven men,
> See me off as I leave the City gate.
> My covered coach is ready to drive away;
> Flutes and strings blend their parting tune.
> Hopes achieved dull the pains of parting;
> Fumes of wine shorten the long road…
> Shod with wings is the horse of him who rides
> On a Spring day the road that leads to home.
> **Po Chu-I 772-846AD After passing his examinations**

The awarding bodies used for the research of this dissertation are accredited by the Office of the Qualifications and Examinations Regulator (Ofqual) and follow the Qualifications and Curriculums Authority (QCA) (2008) GCSE, GCE and AEA Code of practice as their governing document. The professional body used as part of the research, although not offering accredited qualifications, has aspirations to follow the QCA guidance and is currently consulting in order to adopt the approach. In this chapter I will review the literature relating to what affects the reliability of examiners marking based on written public examinations as opposed to classroom assessment.

The literature for this dissertation was researched prior to my sample being interviewed so as to inform my interview schedule and it involved mainly contemporary research. However there are instances where I have used older sources of reference material when I have felt that the observations made and the research undertaken is just as valid now as it was when it was first published.

The outcomes of examinations and in particular public examinations often play pivotal roles in determining the directions that people take at the end of both compulsory schooling and following both further and higher education courses. An example of these in the United Kingdom are the examinations for General Certificates in Secondary Education (GCSEs) which influence whether many thousands of school leavers can proceed to further education or enter into employment. In the current climate competition for ‘good’ schools and university places is fiercer than ever so it is essential that public examinations, and arguably all examinations, are marked as accurately as possible, ensuring fair results for all.

As Suto and Nadas (2008) found, within the broader educational assessment community, it has long been established that when marking public examinations in the UK, inter-marker agreement is imperfect, varying significantly among examination subjects as well as among teams of markers (Valentine, 1932; Murphy, 1978, 1982; Newton, 1996; Pinot de Moira, Massey, Baird and Morrissey, 2002; Laming, 2004).

So fundamentally why do examiners play such an important role in education and what affect can unreliable / aberrant marking have. As Filer and Pollard (2000) assert;

> ”acceptance of outcomes will depend on perceptions of the ‘legitimacy’ of systems of assessment. The concept of ‘legitimacy’ is crucial in this as, throughout history, the outcomes of assessment have been economic and social rewards for some, reduced access to educational and occupational opportunities for many. The mass categorising and social differentiation of populations have needed to be accepted as broadly just, in particular by the loser in the assessment stakes” (p. 128).

The quote supplied by Filer and Pollard (2000) is of special importance within the current climate of press and educational establishments, accusations of examinations becoming easier and it being harder than ever to differentiate and rank students. But more fundamentally examinations do have to be ‘legitimate’ both in their setting and in their marking. The marks received by candidates and their subsequent success or failure in examinations should be appropriate and warranted.

Marking Strategies

First of all it is important to understand the recognised process(es) by which examiners mark. According to Greatorex and Suto (2005) there are five cognitive strategies:

- Matching;
- Scanning;
- Evaluating;
- Scrutinising;
- No response.

Essentially an examiner adopts the matching strategy when:

> ”…the answer to a question is a visually recognisable item or pattern, for example, a letter or part of a diagram. The examiner looks at a particular location in the answer space and judges whether the candidate’s answer in that space matches the mark scheme answer” (Greatorex and Suto, 2005, p.4).

Scanning has been identified in a number of established and well regarded psychological studies for example Kramer, Coles, and Logan. (1996). In essence examiners use it when:

> “…they survey the whole of the answer space designated to a question to find whether a particular detail in the mark scheme is in the candidate’s answer. This detail could be simple, for example a letter or part of a diagram. Alternatively, it could be more complex, for example, a point in an argument; in such cases, further cognitive marking strategies might also be used” (Greatorex and Suto, 2005, p. 4).

Evaluating has been identified where an:

> “…examiner pays attention to either all or part of the answer space for a question, and the candidate’s answer is processed semantically. The examiner awards marks, bearing in mind the structure, clarity, factual accuracy and logic or other characteristics of the candidate’s answer given in the mark scheme” (Greatorex and Suto, 2005, p. 4).

Scrutinising can and does follow on from the above, or is used together with other cognitive strategies but is used only where a response is unpredicted. An examiner:

> “…tries to establish whether the candidate has given a valid alternative to the answer in the mark scheme. To do this, the examiner evaluates numerous features of the candidate’s response with the overall aim of reconstructing the candidate’s line of reasoning or establishing what the candidate has attempted to do” (Greatorex and Suto, 2005, p. 4).

The final strategy is self explanatory. The ‘no response’ strategy is used when a candidate has failed to provide a response to an item(question) in the answer space provided, the examiner looks at the space once or more and then gives 0 marks.

Greatorex and Suto (2005) found that different strategies were used among different examiners; however, in their study they found that the most obvious and prominent differences between marking were between subjects and questions and when marking, examiners tended to use strategy combinations rather than single strategies. They also found that:

> "…no clear relationships between strategy usage and marking reliability were found, suggesting multiple successful ways of marking some questions" (Greatorex and Suto, 2005, p. 5).

Greatorex (2007) predicted that examiners might begin marking a question using a particular cognitive strategy but later in the marking process they might use different cognitive strategies e.g. scanning, when they become familiar (so they believe) with both mark schemes and the responses provided by candidates.

Therefore there appears to be no correct cognitive strategy that should be used to mark a particular question or any correspondence between the method used and the resultant accuracy of its marking. Additionally all of the strategies discussed were found to have been used by both experienced and inexperienced examiners.

Based on what is known about how examiners mark, the strategies can be taught / communicated to examiners but which strategy they use when marking should not be dictated as there appears to be no 'correct' approach.

The importance of consistency of marking

Aslett (2006) found that there are two main forms of examiner reliability:

- Intra; and
- inter- rater reliability.

Intra-rater reliability can be defined as the:

> ”… internal consistency of an individual marker” (Aslett, 2006, p. 86).

Whereas inter-rater reliability is defined as the:

> “…consistency between two or more markers” (Aslett, 2006, p. 86).

There is an argument that intra consistency should be considered the more important of the two as without internal consistency over a series of scripts the marks assigned will be haphazard and unjustifiable and no form of moderation or adjustment of marks will be able to resolve this. In a practical sense this can result in an entire batch of scripts marked by an examiner having to be remarked because a mark adjustment can not accurately be made e.g. + 4 marks. That is that some of the scripts will warrant the extra 4 marks, whereas others will not.

This was explored further by Thyne (1974) who reasoned that although marking-consistency is necessary for maximum validity; *other* conditions also need to be fulfilled. The example given by Thyne (1974) considers two self-consistent examiners who produce different marks on marking, independently, the same set of scripts. If either of these examiners were the sole examiner, his marks would satisfy the condition of marking consistency, however the two sets of marks, in this case, would be different. The valid question is then - can two self-consistent but *different* sets of marks for the same scripts both be valid? This can depend on the following:

- They could be different yet compatible – the two examiners were in complete agreement about the merits of the scripts e.g. candidate ‘A’ being top for both examiner but they may have awarded different marks; and
- The examiners were in complete agreement about the relative ‘distances’ between the merits of each script e.g. candidate ‘A’ was twice as good as candidate ‘B’.

Problems arise however when:

> ”… there was not a perfect correspondence between the two sets of marks, the sets would be incompatible if Charles (candidate A) came out top for one marker but sixth for the other. Obviously, two incompatible (non corresponding) sets of marks for the same scripts cannot both be fulfilling the one purpose” (Thyne, 1974, p.12).

The purpose of this example is to highlight that both examples of examiners marking were self-consistent, and that since at least one of the examples supplied cannot have maximum validity, it is possible for a set of examiners to be marking consistently and yet be invalid:

Pidgeon and Yates (1968) also found that examiners often differ in terms of awarding marks with one examiner finding little or nothing to choose between a given set of scripts and may therefore award all of the scripts the same mark or grade. Another examiner may be more discerning in his / her marking of scripts. He / she may perceive qualities in one that are absent in the other and accordingly mark them differently. A third examiner may perceive distinctions that are over refined and may be pernickety enough to reward or penalise students for what other examiners would be inclined to regard as trivial differences. Another major difference that examiners may betray is disagreement about the relative merits of a set of scripts. Two examiners might adopt broadly similar standards and employ equivalent degrees of discrimination but might nevertheless be disposed to place the same group of pupils in somewhat different rank orders. This could occur if, for example, they disagreed about the importance of various aspects of a syllabus / curriculum and were disposed to react differently to the inclusion of particular kinds of skill of knowledge.

Individual factors influencing reliability

Aslett (2006) found that there are various physiological and psychological variables that affect examiners' reliability. These included:

> "Fatigue, either mental (lack of interest/repetition) or physical (lack of sleep), has been found to significantly affect the reliability of the marks assigned by an individual assessor. Mental fatigue due to monotony and lack of interest in a task can have severe implications with regards to task performance and accuracy" (p. 86).

And:

> "….lack of sleep, whether sleep deprivation or fractal sleep disturbance can lead to lassitude affecting vigilance, attention, logical reasoning, and rational thinking" (p. 87).

Wolfe, Moulder, and Myford, (2001) developed the term Differential Rater Functioning over Time (DRIFT) which was used to describe how the accuracy of a single examiner decreases over time due to fatigue and lack of attentional control. As a result of the DRIFT condition equivalent answers marked earlier by an examiner can be found to receive significantly different marks to answers marked later on. In the study by Klein and El (2003), they also found that papers marked earlier in a marking session were awarded significantly lower marks than later marked papers.

Aslett (2006) also found that emotional factors can play a part in the marks that examiners award. This was demonstrated to be most obvious when examiners were aware of the identity of the student whose work they were marking:

> "Whilst an assessor would hope to remain as objective as possible throughout the assessment process, where a marker is aware of a student's identity, their marking can potentially be profoundly affected" (p. 87).

Research suggests the most common expression of behavioural factors affecting examiner reliability is demonstrated by an examiners stringency and / or leniency.

Spear (1997) found that examiners over mark good work following a poor quality submission and mark harshly when assessing a poor piece of work following a substandard submission, therefore leading to potential intra-rater reliability bias.

Weigle (1998) and Ruth and Murphy (1988) both observed that inexperienced markers were more stringent than experienced assessors thus creating inter-rater reliability bias.

Ecclestone (2001) cited in Aslett (2006) gave the reasons for this discrepancy as unclear; however, possible factors may include novice markers being more “rule –based”, more deliberative, more observant of the assessment criteria and taking more time in their marking.

Ecclestone (2001) cited in Aslett (2006) also found that novice markers could be much more accurate than their experienced counterparts who could place greater importance on their intuition. Ecclestone (2001) suggests the attitudes of experienced markers are imbedded so deeply within the experienced assessor that they are not able to articulate their reasons for assigning a particular mark as their reasoning moves from concrete to abstract over time with increased experience.

Suto and Nádas (2008) generalised that marking could be affected by both (i) the demands of the marking task, including marking strategy complexity, and (ii) a marker’s personal expertise. They further argued that, accuracy can be improved both by reducing the demands of the marking task and by increasing a marker’s personal expertise. **Figure**

1 conceptualises some key factors identified as likely to contribute to marking accuracy, (adapted from Suto and Nádas, 2008)

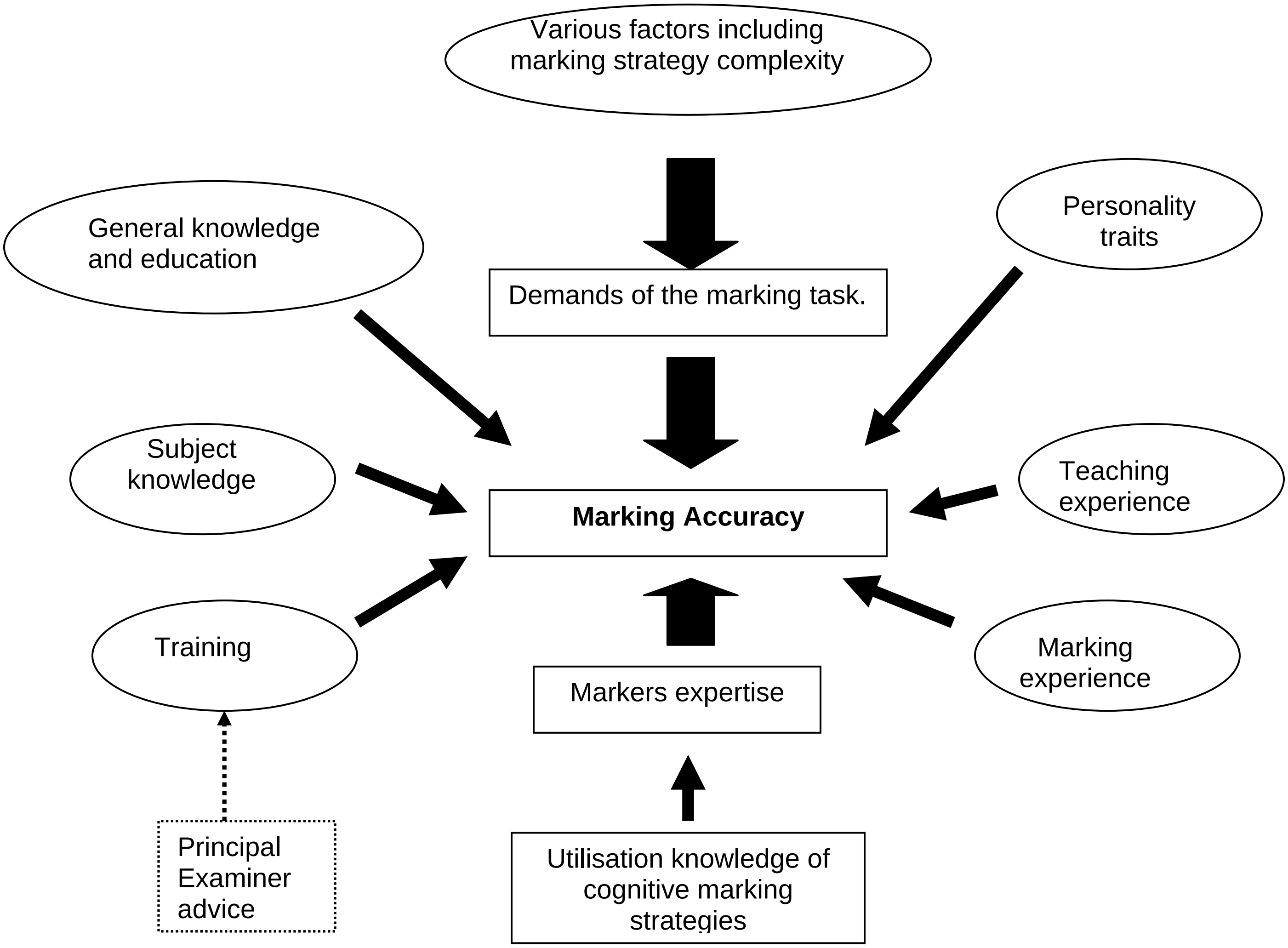
Various factors including marking strategy complexity
General knowledge and education
Demands of the marking task.
Personality traits
Subject knowledge
Marking Accuracy
Teaching experience
Training
Markers expertise
Marking experience
Principal Examiner advice
Utilisation knowledge of cognitive marking strategies

Figure 1

Interestingly, furthermore Suto and Nadas (2008) found that the level of a marker's highest education achievement (either in general or in a relevant subject) is essentially a better predictor of accuracy than either teaching or marking experience. This is of huge relevance to the organisations used as the sample for this study, who use highly qualified and vocationally experienced practitioners rather than teachers and lecturers i.e. individuals who understand the concept of assessment.

In an *earlier* study undertaken by Suto and Nadas (2008), it was found that graduates in relevant subjects but with neither teaching nor marking experience were able to mark as accurately as individuals with both teaching and marking experience. They therefore broadly suggest that when it comes to marking:

> ”…education (of an examiner) is more important than experience” (p.10).

They do however assert that:

> “…suggesting that a marker’s highest level of education in any subject is a better predictor of accuracy than his or her highest level of education in a relevant subject is open to a number of interpretations. The most likely of these is arguably that the key to successful marking is being able to follow marking instructions and interpret the mark scheme in the way its author intended” (p. 10).

In essence, somebody may have a high level of qualification but they still need some form of instruction and training in how to apply, for example, a mark scheme and some degree of aptitude for the role of being an examiner.

Wolf and Silver (1986) cited in Torrance (1996) found that some examiners / assessors demanded perfect performance (albeit on a fairly simple exercise) for a student to be deemed competent, while others were satisfied with performances which fell well short of this standard:

> ”…the assessors behaviour showed a universal tendency to ignore written instructions in favour of their own standards and judgments” (p. 98).

Gender and Marking Reliability

For reasons that were not pursued as part of this study, the marking of public examinations tends to a male dominated pursuit and Greatorex and Bell (2004) undertook a study to discover if gender has any significant influence on marking reliability and found that there was no discernable relationship between the two.

There have been other studies which have focused upon sex bias and gender bias for example. Gipps (1994), found that bias can occur when the overall mark given by an examiner is consciously or unconsciously affected by factors other than the candidates’ actual written responses e.g. sex, ethnic origin, handwriting. Alternatively examiners award marks to answers, which illustrate skills, knowledge and/or values irrelevant to the test but which are valued by the examiners themselves e.g. similar religious beliefs

There is some evidence of sex bias, for example, O' Neill (1985) cited in Greatorex and Bell (2004) found that in teacher's assessment of student work, markers devalued the performance of their own sex. It has also been found that examiner behaviour varies with different groups, such as professional background (of particular relevance in vocational qualifications), subject specialism and gender (Hamp-Lyons, 1990; Vann Lorenz and Meyer, 1991). Greatorex and Bell (2004) allied this to presumably being due to each group having a unique frame of reference.

> ”… as a general rule the sex and gender of examiners and interactions between candidate's sex and examiner sex does not affect the marks that candidates gain at the unit level. In other words although examining is male dominated this has not resulted in a bias against girls or boys in the marking” (p. 11).

They did however feel that as good practice based on the findings of their study and those discussed above, that sex and gender bias in marking is something which should be monitored but that it was unlikely to be significant enough to affect overall grades.

Selection of Examiners

When selecting examiners, as we have learnt, it is appropriate to appoint those who have some subject expertise in what is being examined and who have attained qualifications at an appropriate academic level. In addition to selecting appropriately qualified examiners, limiting the numbers of examiners who are used can also improve reliability:

> “Inter testing reliability for individual administered tests can be increased by restricting the selection of testers to trained persons….all of whom follow a standardised procedure” (Lewis, 1974, p. 88).

All awarding bodies in England, Wales and Northern Ireland offering accredited qualifications have to abide by the Qualifications and Curriculum Authority (QCA) (2008) GCSE, GCE and AEA Code of practice. The code supports restricting examiners marking an examination to a minimum for reasons of validity based on what is being assessed :

> “In the interests of reliable marking and to reduce the scope for variability, the awarding body should ensure that marking is undertaken by the minimum possible number of examiners. In arriving at this minimum number, the awarding body must ensure that the amount of marking allocated to examiners takes account of:
> **i** the nature of the unit/component being assessed
> **ii** the time required to mark candidates’ work”(p. 19)

Training of Examiners

Once examiners have been selected, both good practice and statutory regulation dictates that training should be provided in the correct marking practice for the award / qualification to be examined. For those awarding bodies with accredited qualifications through the auspices of the Office for the Qualifications and Examinations Regulator (OFQUAL) or the Scottish Qualifications Authority (SQA) training of examiners is a mandatory requirement. Paragraph 10 of the statutory regulation of external qualifications (2004) in England, Wales and Northern Ireland, requires awarding bodies to have procedures in place to ensure that their 'associates' have access to appropriate training and guidance.

The Qualifications and Curriculum Authority (QCA) stipulated in their "review of question paper setting and senior examiner training, for GCSE and A levels (2008)", that those responsible for training examiners:

> "…identifies training needs for individuals and groups, organises examiner training programmes and produces centralised training and guidance materials" (p. 9).

In the same report the QCA stated that although training of examiners was an important factor in the quality control process and that training:

> …"can also improve the consistency of examiners' individual marking (intra-rater reliability)" (p. 16).

It could not be used as a stand alone management control:

> …"Training can bring examiners' differences in leniency (interrater reliability) to an acceptable level but it cannot eliminate them" (p.16).

It is obviously a sensible approach to train new examiners so they are fully aware of their duties e.g. how to annotate scripts (ticks, lines, numbers etc) and where they should award marks (based on Blooms taxonomy level descriptors, partial responses provided etc). This training may depend on whether they are:

i first-time examiners, who need training on all aspects of the examining process relevant to their role before marking items;

ii new to the awarding body and require training specific to the awarding body's procedures; and

iii new to the particular unit/component or specification and require training specific to that unit/component or specification.

The QCA also require that during examiners' first marking period, and on subsequent occasions if necessary, they should be allocated a mentor, normally a more senior examiner e.g. a team leader, to provide close support throughout the marking period.

Leadership

It is essential in any marking process that there is the provision of appropriate leadership throughout the examination process, from paper setting through to the appeals procedure. The person selected for this role is usually called a principal or senior examiner. The principal examiner is responsible for the setting of the question paper/task and the standardising of its marking. The Qualifications and Curriculum Authority (QCA) (2008) GCSE, GCE and AEA Code of practice states that the principal examiner must:

- seek to ensure parity of standards across optional questions in the paper and assist the chief examiner in ensuring parity of standards across optional papers;
- monitor the standards of marking of all the examiners for the paper, including, where necessary, any assistant principal examiners and team leaders, and take appropriate steps to ensure accuracy and consistency

As a rule principal examiners:

> ”…are highly experienced in the field of the subject and in the techniques of examining. They will have served several years as an assistant examiner and had their work persistently evaluated. Additionally they have no axe to grind in relation to individual candidates…”(Desforges, 1989, p. 67)

The role of principal examiners and influencing examiner reliability starts with setting and devising appropriate questions and mark schemes however they and their team leaders (experienced examiners used to mentor and oversee a small team of examiners) use their experience to identify any variance between examiners. These are discussed at the standardisation meeting and efforts are made to resolve them.

Mark schemes

Mark schemes are fundamentally used by examiners to guide and inform their decisions throughout the examination process, however:

> ”Like all documents, marking schemes are open to interpretation. Added to that, examiners differ in experience and temperament” (Desforges, 1989, p. 65).

And as we have discussed:

> “Some markers are much more willing to give a candidate the benefit of the doubt than others” (Desforges, 1989, p. 65).

Generally mark schemes can be divided into levels-of-response mark schemes and points-based mark schemes (also known as unit counting, enumeration and count scoring) :

> “Levels-of-response mark schemes specify level descriptors – that is, a description of the kind of answer that will receive a mark from within a given band. For example, a level descriptor might read ‘good understanding across the breadth of the material and some synthesis shown in the answer: 4 to 5 marks’. Points-based mark schemes distinguish between the individual tasks that candidates can do and cannot do, and marks are given according to the tasks that the candidate completes correctly – for example, a mark for each correct label on a biological diagram” (Greatorex and Bell, 2008, p. 334).

Although both levels-of-response and points based mark schemes are widely used Thyne (1974) suggested that points based mark schemes are most likely to aid examiner reliability:

> "…the Unit-Counting method of marking is likely to be the most consistent method, and in so far as the criteria have been properly constituted, also the most relevant" (p. 252).

In order to aid reliability of the marking process, the QCA (2008) GCSE, GCE and AEA Code of practice suggests that mark schemes should:

- include general instructions on marking;
- are clear and designed so that they can be easily and consistently applied;
- allocate marks commensurate with the demands of questions/tasks;
- include the mark allocation for each question/task and part of a question/sub-task, with a more detailed breakdown where necessary;
- include marking instructions for those questions where extended written answers are expected and the quality of written communication used by candidates will be assessed;
- include an indication of the nature and range of responses, appropriate to the subject, likely to be worthy of credit;
- state the acceptable responses to each question/task, or part thereof, with detail that allows marking in a standardised manner; and
- allocate credit for what candidates know, understand and can do.

When designing questions it is important to ensure, in fairness to the candidates, that the questions and the corresponding mark scheme used for an examination relate to the relevant syllabus and are not too peripheral to it, the appropriate command words are selected e.g. outline, explain, state etc, questions are based on the academic level being taught and the course material covered. In relation to how question design can positively affect examiner reliability, the question and mark schemes must be designed:

> ”…in such a way that difference of opinion about candidates’ answers will be reduced to a minimum” (Thyne, p. 178).

The marking schemes for any particular examination should obviously be highly specific to it. For instance, it would be adequate for an examiner to decide to ‘award the mark of 1 for each significant point’ and he would have to make clear what the significant points are. As Thyne (1974) suggests:

> ”The more precisely the relevant performances are described in the marking scheme, the more relevant and consistent the marking is likely to be” (p. 248).

Standardisation

The QCA requires that awarding bodies follow a series of quality procedures to standardise marking. These procedures include coordination meetings more commonly referred to as standardisation meetings.

The purpose of the standardisation meeting, along with other control measures, for example, moderation, is to enable valid and reliable marking. In addition to attending a standardisation meeting the examiners submit a predetermined number of their marked scripts to the principal examiner (or team leader if large numbers of candidates are sitting) who reviews their marking and provides personal feedback to them. If the marking is sufficiently reliable the senior examiner deems that the examiner can continue to mark as before. If the marking is not sufficiently reliable then:

> ”…examiners are required to provide a further sample for review and receive more feedback, and sometimes stronger action is taken depending upon the circumstances” (Greatorex, Bell, 2008, p. 334).

Baird, Greatorex, and Bell (2004) found that examiners considered all aspects of the standardisation process to be important, particularly the mark scheme. The examiners maintained that what is written, and how it is written is very important in enabling them to understand where to award marks.

In the Baird, Greatorex, and Bell (2004) study, the impact of standardisation meetings were investigated, in the study examiners were provided with mark schemes and some examiners were provided with exemplar scripts and given feedback about the marking of those scripts. In the second study, the effects of discussion of the mark scheme were explored: all examiners received mark schemes and exemplar scripts, but some examiners did not attend a standardisation meeting. The study found that neither process (use of exemplar scripts or discussion between examiners) demonstrated an improvement in marking reliability.

However, these findings contradict the research undertaken by the same authors, Greatorex, Baird, and Bell, in 2002. It would therefore appear that although examiners think that the standardisation meetings are valuable because it helps them understand the mark scheme and makes the principal examiner's interpretation of the mark scheme clear, the meetings do not necessarily improve marking reliability. Some examiners also expressed that attending standardisation meetings gave them confidence to know that they were marking appropriately and had the same understanding of the mark schemes as that of their fellow examiners.

Following standardisation, on the rare occasion that examiners marking and interpretations of mark schemes is not improved following subsequent supervision / and / or inappropriate behaviour in the standardisation meeting is displayed e.g. unable to accept the common consensus, they are relieved of their task.

Moderation

It could be argued that no examination is perfectly valid, particularly in respect of marking-consistency and sampling of questions, and that small divergences from the standard should be discounted. In an example given by Thyne (1974) a candidate scoring 49, or even 48, might be allowed to pass if the pass mark was 50%. It is easy to have sympathy for this argument based on how examiners could and do mark, but:

> ”…in order to effect it one should know exactly how inconsistent the marks are, so that one can determine exactly how far below the standard one may go” (Thyne, 1974, p. 110).

This is where moderation is seen as essential. Once the marking process has been completed moderation should be undertaken via peer review, of an examiner’s individual marking. Moderation is the review and ratification of assessments, that is, the judgments given of the value of candidates work:

> “Moderation understood as review is essential to monitor the quality of assessment and to ensure that it is fair, to see that procedures are adhered to, and to check on interpretations – that is, how criteria has been applied to cases” (Torrance, 1996, p. 123).

It is the role of a moderator:

> ”…to take steps to ensure that the eventual results are, as far as possible, fair to all concerned” (Pidgeon and Yates, 1968, p. 98).

Pidgeon and Yates (1968) stated that the principle of moderation is that of safety in numbers. The key being that the more people who agree about the overall grade to be awarded to a particular performance the more valid the assessment is likely to be, especially if the moderators involved in the exercise are chosen because they are demonstrably experienced and trustworthy examiners e.g. principal examiners and team leaders.

As this literature review has already explored moderation is important because, individual examiners, although marking efficiently and consistently, may adopt different standards. Some may be relatively lenient; others severe in their judgments of the merits of a particular performance. Others may simply be marking poorly having left it too late to mark appropriately and have adopted a random marking technique i.e give the appearance that a paper has been marked by randomly applying ticks.

In moderating examiners, the team of moderators need to pay attention therefore to these three attributes:

> ”…the standard of marking; the degree of discrimination; and the extent to which the examiners have conformed to what the moderators regard as an appropriate order of merit” (Pidgeon and Yates, 1968, p. 103).

The critical decision that a moderator has to make in this respect concerns the size of the difference (between marks awarded) that may be tolerated, based on the guidance provided e.g. 3% difference from the mark awarded by the moderator. In other words he must be able to identify the kind of discrepancy that is *statistically* significant. If aberrant marking has been identified there are a number of options for an awarding body to adopt. For example scripts judged to have been marked well below the standard may have to be fully remarked, or just remarked around the pass standard. Sometimes, if remaining discrepancies are slight and consistent:

> ”…they can be corrected statistically – for example, if Mr X always marks a little low, we add two points to all of his scripts (Desforges, 1989, p. 65).

Feedback to and from examiners

In the QCA review of question paper setting and senior examiner training for GCSE and A levels (2008) the adoption of an examiner self-assessment scheme was considered a successful means of obtaining feedback from examiners and helped with the monitoring and evaluation of examiner performance. The various reporting mechanisms adopted help recognise good work, identify possible examiners for promotion opportunities, address any training needs and prevent any re-use of failing examiners.

Baird, Greatorex, and Bell (2004) suggested that marking reliability is purported to be produced by having an effective community of practice i.e.

> “A community of practice is a network of people who have a shared ‘project’ (activity) which they continually renegotiate. They also have a shared repertoire of communal resources – for example,

> tools, routines, artefacts, vocabulary, styles and so on which have been developed over time” (p. 335)

Prior to their research, no studies had been undertaken which attempted to verify the aspects of community of practice that have been observed to produce marking reliability. One of the findings of the study that was found to reduce the transmission of error was:

> ”…markers receiving immediate feedback on their decisions. In testing English as a Foreign Language, Wigglesworth (1993) found some evidence that examiner biases, like task type and rating criterion, were reduced following feedback and that inter-rater reliability improved. The role of feedback to examiners has not been thoroughly investigated in the literature, but feedback from senior examiners to examiners should help reduce marking errors” (Baird, Greatorex, and Bell, 2004, p. 333).

Wolf (1995) also argued that assessor networks or discussion between examiners is needed for reliability.

How can examiner reliability be improved?

Once it has been identified what can affect an examiners reliability it is then possible to put in control measures to select appropriate examiners, positively improve an examiners performance and to identify earlier on in the marking process (pre moderation) that inappropriate marking is / is not occurring:

It is not just the responsibility of an examination board / awarding body to manage examiners, self management by examiners can often pay positive dividends:

> “It goes without saying that taking regular breaks and not marking when already tired are vitally important points to bear in mind. Revisiting earlier marked scripts and reviewing scripts marked at the end of any marking session is also essential. Marking question by question rather than script by script may also reduce some elements of fatigue as it minimizes cognitive load and enables the marker to get into the mindset of the question” (Aslett, 2006, p. 89).

However it is generally the responsibility of the awarding body (and in the context of this study, the professional body too), to do whatever is practicable to ensure that the marking process is robust and valid.

Future Developments

As new technology develops examination boards in particular e.g. the Assessment and Qualifications Alliance (AQA) and Oxford Cambridge and RSA Examinations (OCR), are altering the way in which they require examiners to mark on their behalf. One opportunity is by asking examiners to mark online which facilitates new opportunities for analysing item marks during marking and identifying patterns that might indicate aberrant awarding of marks. With propriety software such as the DRS e-Marker system or ePen, awarded marks may be collected and analysed throughout the marking process, effectively allowing senior examiners / team leaders, to observe marking in real time:

> “The results can be used to alert marking supervisors to possible quality issues earlier than is currently possible, enabling investigations and interventions to be made in a more timely and efficient way” (Bell, Bramley, Claessen and Raikes, 2007, p. 18).

Bell, Bramley, Claessen and Raikes (2007) describe how effectively paper scripts can be scanned and the images transmitted via a secure internet connection to examiners working on a home P.C. Once the marking of digital scripts has been implemented, marking procedures with the following features can be more easily implemented:

- Random allocation: each marker marks a random sample of candidates.
- Item-level marking: scripts are split by item – or by groups of related items – for independent marking by different markers.
- Near-live analysis of item-level marks: item marks can be automatically collected and collated centrally for analysis as marking proceeds.

The huge benefit therefore of online marking over traditional pen and paper marking is to speed up the detection of aberrant marking by directing marking supervisors’ attention to the examiners most likely to be awarding marks inappropriately.

Testing is also currently being undertaken by the parent group of one of the largest examination boards in the United Kingdom – Edexcel, on using computers to assess English tests. It is hoped that the computers will be able to "read" and "mark" test essays, in essence undertake the role of an examiner.

This technology is in its very earlier stages and as quoted on thetimesonline by Bethan Marshall, senior lecturer in English and Education at King's College London, she states just some of the reticence and concern that the introduction of such technology would bring:

> "A computer will never be unreliable. They will always assess in exactly the same way. But you don't get a person reading it and it is people that we write for. If a computer is marking it then we will end up writing for the computer" (accessed Sept 09).

It remains to be seen whether the introduction of such technology will ever improve the reliability of the marking process by making the 'human' element of examining less of a factor in the setting and marking of examinations.

Summary

When considering examinations and marking, it would seem, according to the literature, that human judgement is probably the best method that awarding bodies currently have to offer.

But it is important to remember that examinations, however unpopular, compel candidates:

> ”…not only to acquire knowledge and skills but to reproduce their knowledge and apply their skills” (Pidgeon and Yates, 1968, p. 5).

Pidgeon and Yates (1968) also observed that:

> “It is also useful for a pupil or student to be able to obtain from time to time an objective and independent estimate of his progress and attainments and to be able to compare himself in these respects with his contemporaries. The damage to morale or even to mental health that might result from unfavorable comparisons is often stressed by those who object to examinations, but it may sometimes be in an individuals best interests to discover his true status (p5).

Ultimately marking is a skilled and hugely responsible task with many variables as to how an examiner will perform. Examiners need to be appropriately selected (based on competencies for more specialised and higher level awards), trained and monitored both during and post marking. It is also worth noting that:

> “…ultimately, the level of marking accuracy deemed satisfactory for questions entailing more complex marking strategies is a matter of judgment, given that such questions entail an unquantifiable but inherent degree of subjectivity” (Suto and Nadas, 2008, p. 10)

Methodology

Methodology

The area I have chosen to identify as a topic to be the subject of a research study is relating to what affects the reliability of vocational examiners' marking.

This research is seeking to answer the question 'what affects the reliability of vocational examiners' marking?' and the answers given by the individuals who participate in the study will be their own experiences of this and I am attempting to build conclusions from these experiences and to analyse the different perspectives on it.

In this chapter I outline the methodology that has developed this research from a proposal in to an operating project. I examine the chosen sample, the method of data gathering and the chosen method of analysis.

The Sample

In order to provide this study with the information it required there needed to be a suitable sample selected who could partake in the study, freely giving information and answering the research question based on their own and their organisations experiences of dealing with and managing examiners. There are limitations to the size and location of the sample. Firstly the sample needed to have experience of using subject experts / non-educationalists as examiners. The sample selected needed to have enough experience of examining to be able to give a reasoned judgement on how their organisation managed the examiners it used and the sample selected also needed to be employed in a senior management position so as to carry enough influence within their respective organisations to have had exposure to examiners marking in a practical sense.

The number of participants was limited as the study is of a small scale and using a single researcher, as opposed to a group of researchers; therefore it was restricted in time and resources. However in order to gain various perspectives I anticipated that the study of participants from two different examination boards and one professional body, all recognised and respected in their field of expertise and all offering qualifications on the subject of health, safety and environmental management, the sample selected was a viable and appropriate option.

The selection of the sample for this research was purposive in that as a researcher I “select particular [people] because they are seen as instances that are likely to produce the most valuable data” (Denscombe, 1998). Participants were chosen for their willingness to be open to the process. Both the awarding bodies / professional body and interviewees have had their names changed for reasons of confidentiality.

Pilot study – description of sample

Awarding Body 1 (AB1) was selected for the pilot study. AB1 is one of the world's most recognised occupational health, safety and environmental organisations operating in over 50 countries. It offers professional membership and health safety and environmental consulting in addition to its being an awarding body. Founded in 1957, it has a turnover of more than £9 million per annum. Its awarding body is recognised and accredited with the Office of the Qualifications and Examinations Regulator (Ofqual) which means it has met and must adhere to a wide range of quality assurance criteria so that rigour and consistency in the awarding of qualifications is maintained.

The person selected for interview was AB1’s Director of Awards ‘Alan’. As Director of Awards Alan has overall responsibility for some 80,000 examinations held each year, the majority of which are examined via multiple choice examinations. The multiple choice answer sheets are marked via the use of optic mark readers, so do not require the use of examiners. Approximately 1000 of AB1’s examinations each year are marked by examiners. AB1 offers qualifications ranging from ‘entry level’ to Level 6 diplomas’ and utilises both in house staff and external

‘Senior Examiners’ in the production of examination papers and the marking and monitoring of examination papers.

Post pilot – description of sample

Post pilot Awarding Body 2 (AB2) and Professional Body 3 (PB3) were selected as the sample.

AB2 was formed in 1979 as an independent examining board and awarding body with charitable status. It offers a comprehensive range of globally-recognised, vocationally-related qualifications designed to meet the health, safety, environmental and risk management needs of all places of work. AB2 courses attract around 30,000 candidates annually and are offered by over 400 course providers in 80 countries around the world. Its qualifications are recognised by the relevant professional membership bodies for the safety, health and environmental disciplines, including the Institute of Occupational Safety and Health (IOSH) and the International Institute of Risk and Safety Management (IIRSM).

AB2 examinations and assessments are set by its professionally qualified staff assisted by external examiners; most of whom are Chartered Safety and Health Practitioners operating within industry, the public sector or in enforcement.

In October 2000, AB2 became the first health and safety awarding body to be accredited by the UK regulatory authorities: The Qualifications and Curriculum Authority (QCA) in England (now Ofqual), the Department for Education, Lifelong Learning and Skills (DCELLS) in Wales and the Council for the Curriculum, Examinations and Assessment (CCEA) in Northern Ireland. In addition AB2 is also an ISO:9004 registered organisation.

The person selected for interview was AB2's Standards Manager 'Jane'. As with Alan, as Standards Manager Jane has overall responsibility for the examination process, from ensuring that suitable examinations are set and that the marking process is robust. Of the 60,000 plus assessments held each year (qualifications have multiple examinable units), all of them require the use of examiners to mark them, as to date AB2 offers only qualifications accredited to either level 3 or 6.

PB3 is a not-for-profit membership organisation established to:

> "promote best practice standards in environmental management, auditing and assessment".

With over 14,000 individual and corporate members based in 87 countries, PM3 is now a leading international membership-based organisation dedicated to the promotion of sustainable development, and to the professional development of individuals involved in the environmental profession, whether they be in the public, private or non-governmental sectors. It aims to:

> "…provide recognition, through high-quality professional qualifications, of those individuals who are competent environmental sustainability professionals and to be recognised as a leading organisation in this field."

And:

> "To contribute to the development of skills and competencies of environmental sustainability professionals through training, information and experience exchange, and the sharing of good practice."

PM3 does not run Ofqual accredited courses but is ambitious in this regard and is undertaking development work to standardise its examination processes and to reflect best practice. Its awards are globally recognised and are often a prerequisite for a job in environmental management.

PM3 offers a range of awards from notional level 2 through to notional level 6. It also offers an open book examination to gain associate membership of its organisation.

'Sally' is PM3's Director of Membership Services and has responsibility for all of the awards offered by PM3 and is also responsible for the open book examination. All written examinations are examined externally by subject experts and unlike both AB1 & AB2 require examiners to negatively mark when required e.g. overlong responses.

It is the responsibility of Alan, Jane and Sally to identify and recruit the required number of examiners, moderators, etc for an examination and ensure that all such appointees have the appropriate skills and competencies to carry out their duties. They also have to ensure that the necessary processes are in place to achieve consistency of standards in examination setting, marking and moderation.

The first interview was with Alan (AB1) and was used as my pilot interview and took place on 3rd April 2009 at a meeting room in the interviewee’s office accommodation. Following this pilot I made modifications to the methodology by formulating themes and emailing the future participants in advance so as to focus the interviews. The second interview was with Jane (AB2) and this took place on 29th April 2009 in the boardroom of Jane’s offices. Sally (PB3) was interviewed in her office on 6th May 2009

Using the interview research method means that the environment where the research has taken place is important and conducive to holding an interview. I therefore ensured in advance that the environments were interviews were held were suitable for taping, were appropriate for reasons of confidentiality e.g. were private, and that the interviews were not advertently or inadvertently interrupted.

When undertaking my research I had to consider factors such as the time aspect of the data gathering, the cost of travel to London, Leicester and Lincoln respectively, food and recording equipment, time for transcribing and also the wider considerations such as family, childcare and work. I have three children, who are aged five, two and also a baby who at the time of writing is only a few months old, and this has an impact on the time and energy I have available.

Interview as the Methodology.

Firstly the definition of an interview as proposed by Kvale (1996) is:

> "An interview is literally an inter view, an inter change of views between two persons conversing about a common theme. In post modern thought there is an emphasis on knowledge as interrelational and structural, interwoven in webs of networks." (p44)

This proves to be a comfortable situation for me as a researcher. I am familiar with working on a one to one basis, and although the relationship is that of interviewer and interviewee and not of, say, a manager and client, the conversational quality of an interview requiring the interviewer to have a "sense of good stories to be able to assist the subjects in the unfolding of their narratives" (Kvale, 1996) seemingly fitted well with my skill set and experience. Kvale (cited in McLeod 1994) also suggests that the interview structure should be:

> "Presuppositionless. Rather than coming with ready made categories and schemes of interpretation, there is an openness to new and unexpected phenomena … it is not entirely non-directive, but is focused on certain themes" (p. 81)

A semi structured interview lasting up to just over an hour in one instance (the pilot) but was pared down to approximately half of an hour when interviewing 'Sally' and 'Jane' post pilot, allowed the participants to communicate experiences whilst also giving them freedom to allow other realisations the opportunity to emerge. A benefit of the interview as a technique is that:

> "Questionnaire responses have to be taken at face value, but a response in an interview can be developed and clarified" (Bell, 2005, p. 157)

However this is only as effective if the interviewer has the skills to expand the responses of the interviewee. The use of a pilot interview assisted me in increasing my skills and awareness and adapting my strategy. It also helped me gain a better understanding of the examination processes that two of the organisations operated under.

Being aware of the risk of bias has been a consideration due to my own experiences of involvement with examining, both as an examiner and latterly from working for an examination board. I am aware that I am loyal to my organisation and feel they are appropriate with the approach they take with regard to managing examiners; I had to be conscious of the fact that other organisations may do things differently, that does not make them wrong. Bias can be apparent by the selection of the sample and the interviewer by, facial expressions, tone of voice and body language during the interview itself and also in the way the data is analysed.

The wording of the questions or themes asked in the interview can also affect the answer given by the interviewee. If the questions are biased or leading then this can provide data that is influenced by the views of the interviewer.

When planning an interview it is important, according to Denscombe (1998), to have “some game plan in mind.” With this in mind the themes that were developed for the interviews were formulated by undertaking research of the topic and then developing open ended key themes to give the interview direction but not a set structure. I conducted the interviews in a semi structured style thereby giving the interview fluidity. As the experiences were diverse the use of themes rather than a rigid question and answer style allowed the participants the space to communicate more freely and as discussed earlier, allowing me a better understanding of their organisations examining methodology.

Time at the end of the interview was given to debriefing if required and allow the interviewees the time to ask me any questions about how the examination board I work for manages examiners and how it attempts to positively affect the reliability of its examiners.

The process is subjective and contains ‘interpersonal dynamics’ (Kvale, 1996) and as the interviewer I was aware that the process can be intellectually stimulating for both parties or it could be anxiety provoking, particularly if either party felt that commercial considerations came into play “talking to the enemy” after all, all parties were mindful that we worked for direct competitors.. Kvale believes it is the “interviewer as a person who is the method, the instrument” and this interpersonal dynamic can result in detailed data and should be acknowledged in the analysis. The identity of the interviewer will also influence the conversation, for example a conventional appearance and a neutral attitude is beneficial so as not to impact on the subject. Denscombe (1998) asserts that an interviewer should be aware of any age gaps and educational differences between the interviewer and the subject, and if there is, to be aware of the impact on the interaction. In practice I was aware of this only once when interviewing the Director of Membership Services from PB3, who I felt appeared slightly embarrassed that their examination process was not as mature and comprehensive as she would wish. I also felt this was because in her eyes, she thought I was in some way representing my more ‘established examination board’, rather than by an interviewer writing his dissertation for a Masters Degree.

Ethical Considerations

It is essential that the researcher “seek(s) the highest possible levels of trustworthiness and integrity” (Bond, 2004), ensuring the integrity and openness of the study is a priority. By informing all participants of the nature of the research in writing and making certain that “adequately-informed, full and freely-given consent…[was] obtained prior to their contribution to the research” (Bond, 2004) the risks of causing any harm to the participant were minimised. By agreeing to remove any “personally sensitive information” (Bond, 2004) the subjects are protected, especially as the places of work and names were mentioned in the interviews.

To ensure the integrity and the ethical stance of the research, interviewees gave their informed consent (appendix 2) and were made aware of the whole process. They were informed that the information given by the individual is to be used in a research dissertation and that it will be read by many, and has the possibility of being published. To establish transparency of the research and to engender the quality of trust, the participants were also given the Research Proposal (appendix 3).

I acknowledged that some interviewees may find the experience of being recorded intimidating so this was discussed and agreed prior to the interview and there was an option for the interviewee to opt out of being recorded. I obtained a written agreement signed by both parties to protect both the interviewer and interviewee. I offered the subjects a copy of the interview transcripts (a sample can be seen in appendix 4) and to ensure that confidentiality would be kept, that identifiable data would be removed from them. To ensure transparency subjects were informed that transcripts were kept in a locked cupboard in a locked office, and it was agreed that the digital recordings would be destroyed on completion of the study.

Participants were not asked to choose a pseudonym, and those names given in this study have been chosen by the author. As a researcher I decided to use pseudonyms, to protect the subjects and also the organisations for which they work.

Pilot

I learnt a great deal from undertaking the pilot not least in terms of when to speak and when to listen, as previously mentioned the interview lasted for over an hour. I also felt that there was an element of nervousness on my part as I wanted to present a professional face but in some way I also wanted to, I believe, subconsciously demonstrate to a fellow professional how much I also new about the role of an examiner and how I felt they should be managed. On listening back to the tapped interview I felt that although I provided more prompts than I would have wished, it was gratifying to note that I did not appear to be trying to align the interviewee with any of my own views that I may have held following my review of the published literature. To minimise the risk of bias I had not spoken to the interviewees, pre and post pilot, about my own views and was much more intentional about suspending my own thoughts to allow the interviews to be more open and fluid following the interview with Alan.

Analysis

The transcripts that were created from the interviews are a different form of data to the oral data collected in the interview. The transcribed interview is “frozen in time and abstracted from their base in social interaction” (Kvale, 1996, p. 92) and does not demonstrate the relationship, the body language, the dialect, as demonstrated by Mason (2002):

> "Do not forget that the transcription is always partial partly because it is an inadequate record of non verbal aspects of the interaction" (p. 77).

Kvale (1996) goes a step further:

> "Although produced as an oral discourse, the interview appears in the form of a written text. The transcript is a bastard, it is a hybrid between an oral discourse unfolding over time, face to face, in a lived situation – where what is said is addressed to a specific listener present – and a written text created for a general distant public" (p. 182)

The transcript is also recorded according to the judgement of the person typing it; therefore the transcript may not be a direct representation of the interview and the words that were fluid in the interview are static and open to interpretation. The time needed to transcribe each interview according to Bell (2005) is "at least four hours of work for every hour of interview" (p. 83). However I found the transcribing to be more time consuming than Bell suggests and I underestimated how much resource it would take.

In order to examine the possible method for analysing the transcribed data it is helpful to define the term 'data analysis,' which, according to Bogdan and Taylor (1975) is:

> “A process which entails an effort to formally identify themes and to construct hypotheses (ideas) as they are suggested by the data and an attempt to demonstrate support for those themes and hypotheses” (p. 79)

As this research is based on the personal experience of the subjects my preference was to gather the strands and themes of the data to draw conclusions in order to have some idea of what in practice the sample felt affects the reliability of examiners marking. So as the original interaction does not dissipate I analysed by firstly removing the repetitions and any digression from the question whilst keeping data that was crucial. Then I categorised the data in to themes or sub sections by colour coding each theme in all the interviews. I was then able to chronologically ascertain any similarities, differences and connections in the experience of the subjects. By submersing myself in the data the categories or themes were reduced systematically to produce the findings. I then “refine[d] a set of generalisations that explain the themes and relationships identified in the data” (Denscombe, 1998).

Reflective Researcher Note

I am aware of a shift within me from the naïve researcher who initially started this process in September 2007 to a more realistic researcher at the end of the process in early September 2009. This has been both in my techniques as a researcher to my understanding of what can actually be achieved in a small scale study. I am aware of the battle to ensure the essence of the data is communicated and there has been a feeling of being disloyal to the subjects who gave me so much insight and personal experience. I believe that this is the start of a process of gaining a wider understanding, and becoming a skilled researcher, as opposed to the end.

Findings

Findings

"The purpose of assessment is to rank the candidates" 'Alan'.

The findings are the result of interviews with three participants from two awarding bodies and one professional body, as discussed in the methodology chapter of this dissertation. The study found that the robustness of the examination process varied between the three organisations represented and demonstrates their different approaches to the examination process and management of examiner's with regard to, time, facilitation, experience and core business demands.

A summary of their accreditations, the academic levels of the awards they offer and subjects offered is shown in Table 1: With regard to PB3 any suggested academic levels are purely notional based on the organisation's own judgement, as the awards that they offer are not accredited and have not been judged as such by an appropriate body as discussed previously:

Table 1

Interviewee	Name of organisation	Accredited by OFQUAL / QCA	Academic level of awards offered	Subjects offered
Alan	AB1	Yes	Entry, 1,3 & 6	Health, Safety & environmental management
Jane	AB2	Yes	2, 3 & 6	Health, Safety & environmental management
Sally	PB3	No	3, 4 & 6 (notional)	Environmental management

During analysis of the interviews with ‘Alan’, ‘Jane’ and ‘Sally’ I have identified five key themes:

1. Barriers to and drivers for, consistency of marking;
2. Training and guidance for examiner’s;
3. Examiner attendance at standardisation meetings and what aids the meetings effectiveness in aiding reliability;
4. Post standardisation monitoring of examiner’s
5. Moderation

Barriers to and drivers for, consistency of marking

When the participants considered what affected the reliability of examiner's, each interviewee believed different aspects affected reliability both positively and adversely.

Alan made particular reference to the assignment marking on AB1's level six diploma. Candidates' non adherence to word counts was of particular concern:

> "I think particularly with the assignment marking, where some of these assignments, you know you set a 6,000 word limit or whatever and these candidates are writing 18,000 words, the examiner's do get prickly about you know, the minimum wage almost. You know they start doing a math; they start doing a math, and they say, "Well actually I'm being paid £2 an hour", which is true they are".

Alan felt that crudely, the longer the scripts supplied and the more the marking is based on 'levels of response' the more the possibility in variations in marking. This was especially of concern as examiners could already have different interpretations of mark schemes due to AB1 only offering standardisation meetings for assignments twice a year as they run an 'on demand scheme for assignments', that is, candidates can request an assignment as and when they are ready for one. At any one time examiners could have three different assignment briefs in their possession. The other area of concern for Alan was that as the marking window, including the issue of results, was only six weeks, AB1 did not have the opportunity for marking review meetings and so took the mark given by the examiner: He was especially concerned with the marks around the pass / fail border line:

> "If they had been marked by another examiner I think the variance is obviously more than it would be with a tight point mark level exam".

Alan's passing comment about examiners' relatively low rate of remuneration was also raised by both Jane and Sally in their interviews. Jane acknowledged that examiner's do raise concerns about how much they get paid for the work they do, but felt quite strongly that the majority of examiner's are much more concerned with ensuring fairness of the assessment process and do not undertake the role for any perceived financial benefits. Vanessa supported this by conceding that PB3's examiners fulfil the role for them, not for financial reasons as they:

> "…require people to commit to all stages, to commit time, really free of charge".

Jane felt that reliability could be adversely affected by the timescales allocated for marking purposes. Although AB2 has a marking window of 12 twelve weeks only three of those are set aside for the actual marking of the papers by examiners, the remaining time being devoted to clerical checks, moderation and result panels. Jane felt that if an examiner did not make full use of the time available to them and ultimately rushed the marking, marks would and could be missed.

However to counteract this, AB2 had recently implemented the use of ‘team leaders’ into their marking procedure and that hopefully they will assist in:

> “…weeding out the examiners that are not giving it the due time and attention it deserves”.

Jane did acknowledge the very real time pressures examiners were under, particularly as they have to wait for feedback from team leaders, but expressed that a balance needs to be struck to aid reliability:

> “The balance is the time that they’ve got available to mark, and from a reliability point of view, obviously the fewer examiners’ you have, the better. But set against that, we’ve got to look at how many scripts there are to mark and the number and available marking days the examiners have. So it’s a balance between that and keeping the number of examiner’s as low as possible”.

Interestingly Jane also made reference to those examiners who are particularly diligent and revisit scripts to ensure that they have applied the mark scheme appropriately. She felt that this could both help and hinder reliability.

However, Jane felt the *biggest* barrier to reliability is where examiners are appropriately qualified to mark, but unable to understand what is required of their role in relation to marking:

> “So I think the biggest problem we have that affects reliability is that examiners who are perhaps academically qualified, if they’ve got CMIOSH (are chartered members of the Institute of Occupational Safety and Health) but the confidence exceeds

> thorough understanding of the assessment process….But occasionally, we get people that don't understand the purpose of standardisation, i.e. that everybody leaves that standardisation meeting with a common understanding of the mark scheme and how to apply it".

When discussing barriers to reliability Sally made reference to the number of scripts that her examiners are issued with. The current practice is to give examiners 10 scripts. When pressed, this was not for any reasons of validity / consistency of marking; it was because that was all the majority of their examiner's were prepared to take. On their Associate Open Book Assessment (which is used to gain professional membership of the organisation), this could result in thirty-five examiner's being used three times a year. Sally also raised concern that she felt they gave the examiners too long to mark the scripts (four weeks), but she felt that this was appropriate as effectively the examiners were marking in a:

> "…voluntary capacity. We tend to be -- give them as much time as possible really to make sure they help us in that regard. So, it's to keep them enthusiastic rather than putting pressure on them".

Sally was also concerned that the question and mark schemes PB3 used for its examinations were drafted internally and that there were on occasion consistency issues, particularly with how examiner's could interpret what marks should be awarded for.

In relation to PB3, Sally felt to aid reliability, having fewer examiners marking papers would certainly help. Interestingly she also felt that they could also move away from examiner's marking a whole paper and instead give them just one question to mark from the whole cohort, ergo enabling examiner's to fully understand the mark scheme and how it should be interpreted.

Both AB1 and AB2, although not limiting examiners to just one question to mark, do encourage 'horizontal marking'. As Alan put it;

> "…if you've got 40 scripts, mark all the question ones then mark all the question twos etc. Which they do now, because then you don't get a bee in your bonnet about a candidates handwriting or whatever".

When describing what he meant by getting a 'bee in your bonnet', Alan expressed that on occasion, when marking exams, examiners will sometimes become frustrated with the authors of scripts whose handwriting is poor, and as poor handwriting usually equates to an examiner having to spend more time marking a script, the examiner may be become harsher when awarding marks (this is in addition to marks being missed due to illegible handwriting). By marking one question at a time from the whole batch (horizontally), the examiners have less opportunity to become aggrieved by a candidate's handwriting or approach to the examination paper.

Training and guidance for examiners

Training of examiners elicited very different responses from the participants. AB2 had the most formal training regime of examiners where there is a mandatory requirement for any prospective examiners to attend a day's training workshop. The training workshop is were they gain a greater understanding of the organisation, some background training on assessment procedures e.g. where and how to annotate scripts, meanings of command words (based on Bloom's taxonomy) etc. The majority of the day however is spent replicating a standardisation meeting enabling the prospective examiners the opportunity to understand the process and to actually start to mark and annotate some mock scripts.

The attendance of a Standards Officer (the exam board official responsible for managing the marking of the award and reporting to Jane the Standards Manager) will also allow him / her the opportunity to gain a feel for an individual's ability and aptitude to mark. Additionally Jane felt that the Standards Officer also had the opportunity to discover whether the prospective examiner had both the personal characteristics e.g. ability to accept a common interpretation of a mark scheme, and as discussed earlier, the theoretical knowledge to apply both a mark scheme and understand the assessment process.

When asked how AB2 would ascertain if the training provided had been successful Jane was very adamant in stating that the training *was* successful. Until three years ago no training had been provided for new examiners other than, at best, mentoring "a little bit by the principal examiner's".

Other indicators as to the success of the training has been that the number of scripts requiring a full re-mark, post moderation, has significantly reduced, although in part this could also be an indicator of the success of the team leader approach and examiners being aware that they are being monitored more closely. Another indicator given was that the number of successful enquiries about results (candidate examination result appeals) had reduced with:

"…much less differences between original examiner marking and the re-mark by the team leader or the principal".

AB1's approach to training examiners was not as formal as that of AB2. Basically AB1 give their new examiners a:

> "sort of 45 minute induction prior to the standardisation meeting. But that really is, "This is what we mean by point marking, this is what we mean by levels of response marking, this is the colour pen you use". It's the good old -- it's the good old examiner briefing.

In essence:

> “A little examiner coming in would receive training on the mechanics of marking, shall we say, rather than anything deep and meaningful”.

The phraseology in this quote used by Alan, is emotive i.e. “little examiner”, and could portray his real view of examiners.

PB3’s approach to marking was again informal. Prospective new examiners are asked to mark a trial paper from a past sitting. Providing the marking looks “in line”, the new examiner shadow marks an experienced examiner during the next round of examinations. Both the new and experienced examiners mark a set of scripts and then come to an agreed set of marks.

As part of the training provided by the organisations, the provision of written guidance was explored with the interviewees. AB1 relied upon the information provided within its ‘specifications’ i.e. syllabi, where the command words are defined e.g. outline, explain, identify etc. Examiners were expected to use the command word descriptors to benchmark how much detail was expected in a response, and then from an answer provided award marks according to a mark scheme. To support this published guidance the advice given by Alan and his team to a newly appointed examiner would be:

> “the same as we would to a candidate at an examination technique day, we say, "Look you know given the time constraint you’re under, an outline question worth five marks is basically, this amount of A4".

Alan felt very strongly that as “the better candidates” adhere religiously to one side of A4 for every 10 marks, or similar, the message has to be consistent:

> “to our tutors, to our candidates and therefore to our examiner’s. You can't be telling the examiner’s one thing and telling the candidates something else”.

AB2 do not currently issue guidance to examiners on marking but coincidently they were planning on releasing written guidance within a few weeks of the interview taking place. The reasons given for this were, because historically, different qualifications had been developed and subsequently managed differently in terms of how and where to award marks. Another reason given was that some of the longer serving examiners had marked under previous management regimes and had adopted different marking practices e.g. marking in the text as opposed to in margins, which makes clerical checking difficult to undertake. It is planned the proposed guidance captures:

> “…all the different rules for marking, where to put the ticks for example. Whether or not to award half marks. All of those issues that are constantly debated, will be captured and become the key document for the Principal Examiner and for the Standards Officers. But it will also be issued at the training workshops and then retrospectively to every existing Examiner on our books and that in itself will be a huge contribution, I think, to reliability.”

As part of the guidance document Jane had “trawled” through as much published guidance as possible e.g. the Joint Qualification Council, the

Federation of Awarding Bodies (FAB), ensuring that when required, it complies with the guidance published by OFQUAL.

PB3 did have rules for marking which are issued to examiners but this concentrated more on the regulations associated with their examinations. For example their Open Assessment qualification has strict word count requirements, failure to adhere to a word count i.e. 300 words per question, plus 10%, results in an automatic failure of that question. In addition any candidates failing to gain more than half marks on three of the ten compulsory questions are automatically referred, and obviously as with any qualification there is penalisation for plagiarism. They do not issue as such, guidance on actually *how* to mark.

Examiner attendance at standardisation meetings and what aids the meetings effectiveness in aiding reliability

Both AB1 and AB2 contractually require examiner's to attend a standardisation meeting prior to marking, failure to attend a standardisation meeting, which are used to gain a common understanding and interpretation of questions by examiners, is viewed to be so essential that non attendance at standardisation results in examiners being prohibited from marking. AB1 and AB2 do however approach standardisation differently.

AB1 carried out something called a 'pre-standardisation' meeting shortly after its examinations have been sat. Alan holds a meeting with the relevant principal examiner and applies the provisional mark schemes to a sample of completed scripts. Alan found this meeting beneficial because:

> "…everything's working fine and you get through to question 4B and you realise by perhaps looking at six or seven scripts that they are coming from a different area of play than you expected. And then it's for us to decide, behind closed doors, you know, the direction that we will give the examiner, examiner's in the standardisation meeting. So I do keep those things quite separate. Meanwhile the examiner's are trial marking at home".

AB1 issues its examiners with five photocopied scripts; in essence all of the examiner's receiving the same five scripts. Alan expressed that he had tried issuing examiners with different live scripts but found that the standardisation meeting becomes "ambushed" when examiner's start asking questions about the papers they have been issued with. Equally he found that some examiners took the opportunity to become too vocal during the course of a standardisation meeting and that hearing:

> "…some examiner on his high horse showing off his knowledge about this, that and the other is not helpful and it's just not very interesting".

Alan therefore found that the amendments made in the 'pre-standardisation' meeting usually cover most of the questions that examiners bring to the formal standardisation meeting. However there was an acknowledgement that examiner’s could still contribute:

> “…we give them a chance to contribute, it’s motivating for them isn’t it to feel they can influence things”.

Although Alan and the principal examiner have usually agreed what they are prepared to move on in, terms of the mark schemes and their interpretations of it, he is prepared to “give a bit”, to aid discussion and as discussed above, motivate.

One of the reasons given by Alan for the restrictive nature of these standardisation meetings is that he would not wish mark schemes to become too long and over generous, accepting that:

> “if you're not careful your mark schemes can, you know -- I think if your mark schemes get too long it's the sign of a poorly worked question to be honest”.

Alan did acknowledge that if you can encourage examiners to be vocal in a managed way the principal examiner and awarding body are able to see whether the directions given are being assimilated or not. Alan felt that another indicator of understanding was whether examiners were making annotations on their mark schemes to refer back to, something he liked to see.

AB2’s approach to the standardisation meetings were less restrictive and encouraged debate. All attending examiners are issued with ‘live’ scripts and at the meeting all questions and mark schemes are discussed giving examiners the opportunity to:

> “raise any concerns or points for clarifications or additions to the mark schemes”.

Indeed those examiners who were not vocal (within reason) sometimes gave Jane and her team concern as to how much understanding they were gaining from the meeting and this lack of participation was sometimes reflected in their marking.

Jane felt as Alan did that the reverse of this behaviour is those examiner’s who are too vocal:

> “that can perhaps indicate, that they’re becoming too precious about their particular points, rather than them fully understanding the standardisation process again”.

Once the mark scheme has been standardised all examiners mark a common script which is then discussed in detail in terms of the spread of marks. Where there are huge differences in opinion the Principal Examiner’s, with the advice and support of the Standards Officer, give the final point of clarification and understanding.

Although PB3 do not hold standardisation meetings they do operate a standardisation process. This is done by issuing the examiner cohort with three common scripts; the scripts are marked and returned to PB3. PB3 can then see those examiner’s who are marking outside of the acceptable thresholds before issuing the ‘live’ papers. Feedback is given and examiner’s informed to adjust their marking if required and they are also provided with a table of their marking against the marking awarded by the responsible person with in PB3 i.e. Sally or one of her team (PB3 do not have a principal examiner role as such but are looking at introducing the role as part of ongoing development work). The reason for issuing the table was:

> “…to justify the mark awarded. So they can see why we've awarded a three or a five or whatever it might be for each of the three papers they've standardised”.

During the marking of the three scripts examiners do have the opportunity to raise concerns if they feel that a question could have different interpretations and mark schemes will be adjusted if necessary. Although done retrospectively PB3 do have an annual meeting with all of their examiners where discussions are held on the:

> “exams undertaken in the previous year, feedback on question style, content, depth, those types of issues and any consistency issues that may have arisen about the marking regime. So we really do open it up and look at everything”.

If an examiner takes issue with the mark feedback given, Sally is alerted to the examiner’s concern and gives direct feedback to them, question by question, and then arranges for them to be moderated more closely once scripts have been marked and returned. Failure to adhere to the common understanding results in direct action:

> “And if there was no improvement, if they didn't then come into line with broadly our marking then we'd ask them to step down”.

When asked what aids an effective standardisation meeting Alan felt that preparation on behalf of the principal examiner was imperative both for credibility and understanding of the requirements of the examining body. Alan also felt that the principal examiner had to be a strong chairman who was able to give high level direction and know when the debating of a point had gone on for too long:

> “…there has to come a point where you have to say, "No, I'm the principal, let it go, accept it and move on".

He also expressed that:

> “I believe if you don't direct those proceedings then, particularly I think in our sector, the creaks will come out the ship. I think if you just go into, you know procedures mode, leave the examiner’s at meetings to do their own thing, they will come off the rails, even the good ones”.

He also felt group dynamics play a part in getting a common agreement on a mark scheme particularly when two examiners are in disagreement and acknowledged that you’ve just “got to let it happen” in a controlled way.

Jane also felt that having an effective and knowledgeable principal examiner chairing a meeting was imperative both in terms of gaining an agreed understanding of the technical aspects of the mark scheme and at the same time being skilled enough to control a meeting whilst allowing time for debate. It was felt that standardisation meetings can be:

> "… influenced very largely by one or two individuals who are perhaps more vocal, and don't always understand the standardisation process. And they become very keen to get their opinion, their addition to the mark scheme, so that is for the Chair to bring it back to the purpose of the meeting. So, yeah, I think pretty much top of the list is an effective Chair, and to keep getting that message across what standardisation is all about. It's not about personal opinions as such, it's literally the standardisation, and the agreement that everybody in the room knows how to apply that mark scheme".

Jane also felt that the facilities provided for the standardisation needs to be conducive to holding a meeting in terms of comfort (both seating and temperature), space and lighting.

Interestingly Jane also made reference to giving examiners assurances during the course of the meeting that any checking or moderation of their marking was in everybody's best interest as some examiners had raised concerns about feeling intimidated by this and it affecting their marking judgements. It was felt that, it was the responsibility of the chair, to reassure examiners that supervision and moderation was of huge importance relating to:

> “increasing reliability which in turn is all about fair assessment, and fairness to the candidates”.

Post standardisation monitoring of examiners

The process AB1 adopts commences with the examiners completing five live scripts, which are then sent to the principal examiner or a team leader if large numbers are sitting, for over marking i.e. remarking of a script already marked by an examiner. The scripts are then returned to the examiners for them to recognise if there were any discrepancies in marking. The principal / team leaders will also give feedback on the marking both written and verbal:

> “The principle examiner picks up the phone and has a nice friendly conversation, so he says, "Look get your marking scheme out. Do you remember what we were saying about question six? Do you remember the point I made in the meeting that you didn't write down?" And typed up that way”.

When asked about ongoing monitoring of examiners Jane responded that AB2 also require examiner’s to mark three to five live scripts post standardisation and forward them to their team leader for over marking. Examiners are advised not to commence full marking of the rest of their allocated scripts until they have received feedback. The feedback will highlight areas of weakness (sometimes due to an examiner’s technical understanding of part of the syllabus in relation to a question e.g. failure tracing methods) or it may be that they are marking particularly harshly or leniently. It may however be that the examiners just require refocusing on one particular question where marking anomalies have been identified. This occasionally comes to light when there is:

> "…an examiner with a specific subject expertise, their expectation can be higher than the level of the qualification on a specific question relating to their field".

Once feedback is given, if the difference between the examiner's set of marks and the team leader's set of marks are significantly different, some additional scripts may be requested to be over marked. If these are then not judged to be appropriate the examiner's batch will be recalled and forwarded to another examiner for marking.

Jane felt that selecting team leaders with both interpersonal and excellent marking skills was an essential part of the process. It was considered essential both in monitoring the marking process, and when required, positively affecting an examiner's mind set when approaching the scripts:

> "The team leaders are, we select based on their historic reliability of marking, so there will always be people who have been examiner's on that particular qualification for at least two years, ideally, and have been able to demonstrate consistent application of the mark scheme, and give appropriate contribution to the

> meetings. That they are able to demonstrate they have got the skills to communicate with the examiner's. This can be quite sensitive when people are working with their peers, that they're actually been advised by somebody else that the markings not appropriate so it is very important they have good communication skills".

PB3 do not undertake any monitoring of examiners once marking has progressed, based in part on the relatively small number of scripts allocated to each examiner. They do however remark any scripts which have been considered to have been marked "borderline". An example of this would be if PB3 had a pass mark of 50% they would remark any scripts in the range 45% to 54%. The reason given for this was:

> "we get those verified to make sure we are awarding the correct mark".

Although the borderline range is generous, it would not necessarily pick up those examiner's who are routinely marking above or below this range.

Moderation

AB1 holds a marking / borderline review meeting. At the meeting Alan, the principal examiner and the team leaders (if used) will undertake a second phase sampling of the examiners' marking. Alan felt this was essential to gain a better overall view of how the examiners had performed. He acknowledged that it was “quite nice” for examiners to be able to select the five papers which they forward for first sampling, chosen perhaps for ease of marking and clarity of answer. But:

> “…the converse is that with second phase sampling the examiner doesn't know which five scripts we're going to pick from their allocation. It’s these scripts that give us a truer reflection on an examiner’s performance and can alert us as to whether or not the examiner is getting lazy”.

When moderating the examiners’ scripts AB1 takes a look at a range of marks i.e. pass, credit, distinction for the purpose of;

> “…informing us how much intolerance is still in the system”.

And equally once they are aware of how an examiner has performed they can use this in:

> “…informing us as to whether or not to offer them a contract for the next session as well”.

Once they are aware of how much ‘intolerance’ is still in the system, which could be for a number of reasons in addition to an examiner’s performance but affected by:

> “…it could be just that one or two questions were a bugger to mark, or the candidates were all over the shop”.

They can use this information to inform them as to how far their remarking scope should be. Alan explained that if they set a pass mark at 57, they would look at all the scripts down to 54. When prompted about those candidates that fell outside of this range, both being under marked and over marked, he felt that so long as first and second sampling of scripts had been done appropriately then candidates will not have been unfairly advantaged / disadvantaged.

This comment in part was based on Alan's perceived knowledge of his examiners. Alan expressed that he would suggest that his accepted intolerance of examiner's marking would be 2% and stand by that judgement if asked by the 'regulators'. That is 2% outside of where a correct and accurate mark would be. If the examining in reality fell outside of the 2% tolerance Alan expressed:

> "And if it gets wider than 2%, remember the context, I've got guys who've been marking together so long they're almost telepathic, then that's fine. Because my borderline review more than covers the 2%".

AB2 undertake a formal marking review involving the Standards Officer responsible for the (relevant) award, the principal examiner and associated team leaders. In this meeting all examiners are moderated and scripts looked at on and around the pass, credit, distinction mark. The team leaders also come to the meeting with an informed decision as to whose scripts they may wish to look at more closely based on the examiner's scripts they have previously marked and may have had some concerns about. Jane and her team set a 5% tolerance on where a team leader has marked a script compared to where an examiner has. Jane explained that:

> “If we are uncomfortable with the standard of marking, within the tolerances we set, we will call for that whole batch to be re-marked., and appropriate feedback will go to the examiner”.

If examiners are found to be within the acceptable 5% boundary, slight statistical adjustment will be given to the examiner’s batch of scripts e.g. all scripts receive an adjustment of plus / minus 2, or it may be that only one optional question will need an adjustment just for those candidates attempting it.

PB3 do not undertake any formal moderation of their examiners, accepting the mark given. The only time this may change is when an examiner has been identified as potentially marking aberrantly during the standardisation process; their scripts would subsequently be remarked by a consistent examiner. Sally also explained that they do adjust an examiner’s marking but do not set a tolerance threshold unlike AB 1 and AB2.

Discussion

Discussion

The discussion chapter of this dissertation includes a brief overview of the study, but the majority of the chapter will be devoted to a summary of the five main themes and associated threads identified in the findings chapter and a discussion of the pertinence of the results on what affects the reliability of vocational examiners marking.

Summary of the Study and Methodology

The outcomes of examinations and in particular public examinations often play pivotal roles in determining the directions that people take at the end of both compulsory schooling and following both further and higher education courses. As numerous studies have found, within the broader educational assessment community it has long been established that when marking public examinations in the UK, inter-marker agreement is imperfect, varying significantly among examination subjects as well as among teams of markers (Suto and Nadas, 2008, Valentine, 1932; Murphy, 1978, 1982; Newton, 1996; Pinot de Moira, Massey, Baird and Morrissey, 2002; Laming, 2004).

Fundamentally examiners play such an important role in education because qualifications should be meaningful and reflect the responses provided by students accurately. As Filer and Pollard (2000) identified, unreliable and/ or aberrant marking directly affect the ‘legitimacy’ of systems of assessment. The concept of ‘legitimacy’ is crucial as the outcomes of assessment can mean economic and social rewards for some, reduced access to educational and occupational opportunities for others.

The purpose of this study was to identify what affects the reliability of vocational examiners' marking; with 'vocational' being interpreted as subject experts rather than educationalists. The interest in this area was prompted during my time spent as both an examiner and as a Standards Officer for an awarding body specialising in vocational awards. It became apparent that there were variances in how examiners approached the marking process and therefore how they marked candidates work. Although there has been extensive research into what affects marking e.g. increased monitoring, clearly structured mark schemes, little research has been undertaken as to what awarding bodies using 'vocational' subject experts have found to be most effective in improving the reliability of the examiners they use.

The assumption of this study was that vocational examiners would need managing differently in terms of selection, training, support and moderation so as to affect the reliability of their marking. What became evident was that the examiners used by the participants in this study needed no more or less management than *any* examiner, working for *any* examination board or professional organisation offering public examinations, regardless of their background i.e. experienced educationalists. This was demonstrated when a review of the literature was undertaken and it was identified that whenever examiners are being used and from whatever field e.g. experienced examiners, teachers, lecturers etc, they are all fallible and that they need support, guidance and monitoring, in essence, using the same practices and procedures as the participants of this study.

As previously discussed, the literature review for this study was undertaken prior to the interviews taking place so as to best inform the semi structured interview question set. The literature researched identified the theoretical and broader overriding issues relating to examiner reliability e.g. educational achievements. *However*, what the interviews identified were some of the practical issues relating to examiner reliability and how these issues can be addressed in a practical managerial sense.

The study found that the robustness of the examination process varied between the three organisations represented and demonstrates their different approaches to the examination process and the management of examiners with regard to, time, facilitation, experience and core business demands.

The study was conducted in two phases. For the first - pilot - phase Awarding Body 1 (AB1) was selected.

AB1 is one of the world's most recognised occupational health, safety and environmental organisations operating in over 50 countries. It offers professional membership and health safety and environmental consulting in addition to its being an awarding body. AB1 is recognised and accredited with the Office of the Qualifications and Examinations Regulator (OFQUAL) which means it has met and must adhere to a wide range of quality assurance criteria so that rigour and consistency in the awarding of qualifications is maintained. The person selected for interview was AB1’s Director of Awards ‘Alan’. As Director of Awards Alan has overall responsibility for some 80,000 examinations held each year.

Following the pilot phase, the questions used in the semi-structured interview were only marginally developed with the majority of the changes being in relation to the interviewers style e.g. a more relaxed approach and knowing when to speak and when to listen.

Post pilot Awarding Body 2 (AB2) and Professional Body 3 (PB3) were selected as the sample.

AB2 was formed in 1979 as an independent examining board and awarding body with charitable status. It offers a comprehensive range of globally-recognised, vocationally-related qualifications designed to meet the health, safety, environmental and risk management needs of all places of work. AB2 courses attract around 30,000 candidates annually and are offered by over 400 course providers in 80 countries around the world. In October 2000, AB2 became the first health and safety awarding body to be accredited by the UK regulatory authorities: The Qualifications and Curriculum Authority (QCA) in England (now OFQUAL), the Department for Education, Lifelong Learning and Skills (DCELLS) in Wales and the Council for the Curriculum, Examinations and Assessment (CCEA) in Northern Ireland.

The person selected for interview was AB2's Standards Manager 'Jane'. As with Alan, as Standards Manager Jane has overall responsibility for the examination process, from ensuring that suitable examinations are set and that the marking process is robust. Of the 60,000 plus assessments held each year (qualifications have multiple examinable units), all of them require the use of examiners to mark them, as to date AB2 offers only qualifications accredited to either level 3 or 6.

PB3 is a not-for-profit membership organisation established to:

> "promote best practice standards in environmental management, auditing and assessment".

PM3 does not run OFQUAL accredited courses but is ambitious in this regard and is undertaking development work to standardise its examination processes and to reflect best practice. Its awards are globally recognised and are often a prerequisite for a job in environmental management. PM3 offers a range of awards from notional level 2 through to notional level 6. It also offers an open book examination to gain associate membership of its organisation.

'Sally' is PM3's Director of Membership Services and has responsibility for all of the awards offered by PM3 and is also responsible for the open book examination. All written examinations are examined externally by subject experts and unlike both AB1 & 2 require examiners to negatively mark when required e.g. overlong responses.

Themes

1. Barriers to and drivers for consistency of marking.
2. Training and guidance for examiners.
3. Examiner attendance at standardisation meetings and what aids the meetings effectiveness in aiding reliability.
4. Post standardisation monitoring of examiners.
5. Moderation.

The five main themes identified above form the basis of the following discussion, however, although I attempt to discuss them in order, there may be some overlap. I will then examine the limitations and strengths of the study.

In the next section the main conclusions for each of the five major themes of the study are reviewed.

Barriers to and drivers for consistency of marking

The participants involved in the study, ‘Alan’, ‘Jane’ and ‘Sally’ were either managing their examiners inline with the protocols set by OFQUAL in their document Qualifications and Curriculum Authority (QCA) (2008) GCSE, GCE and AEA Code of practice or were working towards it. Consequently they were aware of the common and expected approaches (by the regulator) to positively affect inter-rater reliability i.e. consistency between two or more markers by the introduction / use of e.g. team leaders, moderation and clerical checking.

The barriers to accurate marking identified in the literature review found that various physiological and psychological variables affect an examiner’s reliability. These variables included fatigue, both mental and physical which understandably can lead to and has been found to *significantly* affect the reliability of an examiner leading to:

> “…lassitude affecting vigilance, attention, logical reasoning, and rational thinking” (Aslett, 2006, p. 87)

The literature review also discussed the DRIFT phenomenon (Differential Rater Functioning over Time) which was used to describe how the accuracy of a single examiner decreases over time due to fatigue and lack of attentional control. In the study by Klein & El (2003), they also found that papers marked earlier in a marking session were awarded significantly lower marks than later marked papers.

Interestingly none of the participants made reference to fatigue, mental health etc when discussing barriers to accurate marking but majored on the practicalities of the subject based on their own experiences of managing both examiners and awards. Both Alan and Jane made reference to the numeration paid to examiners as causing potential issues in terms of the accuracy of an examiners marking. This was particularly evident when scripts take longer to mark due to a student’s / candidate’s poor hand writing or non adherence to recommended word counts e.g. provide 16000 word assignments as opposed to the 8000 stipulated in the brief. This last point also raises additional concerns, when the marking is based on ‘levels of response’ where there is more opportunity for a variation in marking. All participants felt however that the majority of examiners were not undertaking the role for any financial benefits but alluded to them wanting to “give something back” (to the professions that they represent – health, safety and environmental management) and for reasons of continual professional development.

Only Jane made reference to the length of time examiners have to mark as potentially causing aberrant marking, due to rushing the marking process. Interestingly this is indirectly caused by using as few as examiners as possible as discussed by Lewis (1974) to aid inter testing reliability, which consequentially puts additional pressure on examiners resulting in fatigue (see above) and stress. OFQUAL also requires examination boards to use the minimum possible number of examiners to reduce the scope of variability. Extending marking windows is not usually a viable option for examination boards as timetables have been published and deadlines set. Students / candidates are understandably keen to receive their results and there are also commercial pressures to release results in an accurate and competitive manner e.g. comparable to competitors’ timeframes of releasing results.

However, Jane felt the *biggest* barrier to reliability is where examiners are appropriately qualified to mark, but unable to understand what is required of their role in relation to marking, in essence the confidence of a particular examiner exceeds thorough understanding of the assessment process. This significantly contradicts the Suto and Nadas (2008) study, were it was found that graduates in relevant subjects but with neither teaching nor marking experience were able to mark as accurately as individuals with both teaching and marking experience and that the level of a marker's highest education achievement (either in general or in a relevant subject) is essentially a better predictor of accuracy than either teaching or marking experience.

Sally raised concerns about the mark schemes PB3 use as being far too open to interpretation by examiners and that internally to the organisation there is not currently an appropriate procedure in place to ensure of a consistent quality in terms of the guidance given on them.

In relation to drivers towards accuracy both Alan and Jane felt that marking horizontally e.g. by marking all of the question 1's from a whole batch of scripts before moving on to mark all of the question 2's, that the examiners have less opportunity to become aggrieved by a candidate's handwriting or approach to the examination paper and as a result do not tend to get frustrated with a particular script. This is an important observation because frustration by an examiner may result, either consciously or subconsciously, in harsher marking by an examiner. The reverse of this is that an examiner can also start to will somebody to do well if they take a particular liking to a script and a students approach to an examination. An additional benefit of horizontal marking is that examiners become very familiar with mark schemes and their interpretation. Sally was also considering taking this one stage further by issuing her examiners with just one question to mark from an examination e.g. 15 responses for question 1, which could be targeted based on an examiners specialisms / subject knowledge.

Horizontal marking was also recognised by Aslett (2006) in that by marking question by question rather than script by script it also reduces some elements of fatigue as it minimises cognitive load.

None of the participants made reference to any self regulation by examiners as stipulated in the literature review e.g. not marking when tired, good planning etc. This may have been due to its fundamental nature and the participant's expectation that this element of self policing should not need to be overtly brought to examiners attention. It would be prudent to include any such observations in any guidance supplied to examiners so as to make them fully aware of the physical effects of fatigue etc on performance.

Training and guidance for examiners

Training of examiners has been researched extensively and has been found to be effective in bringing inter-rater reliability to an acceptable level although not eliminate it altogether.

Training examiners in different approaches to marking can be effective as examiners tend to use strategy combinations when marking e.g. matching, scanning and evaluating, because it has been found that when marking, examiners tend to use strategy combinations rather than single strategies. However it has also been found (Greatorex and Suto 2005) that there is no clear relationship between marking strategy and marking reliability which suggests multiple successful ways of marking some questions.

Of all of those interviewed only AB2 had any formal training procedures for its examiners where there is a mandatory requirement for any prospective examiners to attend a days training workshop where they learn how to annotate scripts and gain a better understanding of both the meanings of command words (based on Blooms taxonomy) and the type of responses they should elicit from candidates. It is during this meeting that a member of AB2's 'Standards Department' has an opportunity to see if a prospective candidate can not only mark accurately but also has the personal attributes to be an effective examiner. As Jane alluded to throughout her interview, she felt strongly that examiners do not only need professional and theoretical subject expertise coupled with the ability to understand the assessment process, they also have to have the personal characteristics to be able to accept a common interpretation of a mark scheme, even if this differs from their own way of thinking and to mark with that in mind. Alan also made reference to examiners' personal characteristics as being something that can cause conflict during the examination process in relation to the standardisation meeting and being outspoken. Obviously the correct selection of examiner at the outset can pre-empt this conflict e.g. the ability accept feedback, reflective, prepared to contribute to the debate but also willing to accept the common / group ruling and apply it.

AB1 did not undertake any formal training as such for examiners, which is surprising as it is stipulated as a requirement in the OFQUAL guidance, although the organisation did give a brief overview of the examination process or as Alan referred to it the "mechanics of marking" e.g. what colour pen to use. Alan also spoke about informing examiners verbally about the amount of text he would expect in any given answer, dependant assuming on the command word used in the question e.g. explain, describe, list etc. The 'guidance' given by Alan is obviously very open to interpretation and does not consider what may or may not be actually contained within the answer.

PB3 undertook a much more ad hoc but semi structured approach to training its examiners and did, as will be discussed later, monitor them once marking had commenced. It also protected students from new examiners in the first instance by requiring prospective examiners to shadow mark existing reliable markers.

AB2 were able to demonstrate that the training and the associated cost of that training had been successful in improving the reliability of its examiners and this was demonstrable in the reduction in the number of candidate appeals that had been successful.

Guidance

All three of the sample produces well considered and thorough syllabi / specifications which approximately set out the:

- Structure and rationale of the qualification / award;
- Assessment and criteria methods;
- Full syllabus / specification content.

The syllabi do not set out instructions (e.g. detailed mark schemes) for examiners other than for assessments that are undertaken by external examiners e.g. workplace assessments. Alan spoke about AB1 examiners having access to their specifications which do contain command word descriptors. It is essential that candidates pay attention to command words (also known as action verbs) as they will lose marks if a question asks for an 'outline' and only a list is given. Alan's argument was that these descriptors can guide examiners as to how and when to award marks, however the descriptors are very open to interpretation by examiners, even those experienced at the task.

AB2 were keen to produce, and indeed were in the process of doing so at the time of the interviews, some written guidance which will become the key reference document for all of those involved in writing and marking its examinations and Jane felt by having this clear guidance it would be a "huge contribution, …, to reliability". It is intended the document covers not only the administration requirements and rules of examining but also some mock answers that have been marked reflecting the command words. These examples can then be used as a source of reference and also to dispel any myths e.g. the use of half marks for partial answers.

It can be ascertained therefore that in order to affect reliable marking it would be prudent to issue guidance clearly stating the rules for examining and when and where to give marks. Guidance can be found in numerous reference documents for example those provided by OFQUAL, FAB and the JQC. Although it is advised that examiners need clarification as to when marks can be awarded e.g. when partial answers are provided, as Wolf and Silver (1986) found, assessors do sometimes show a tendency to ignore written instructions in favour of their own standards and judgments, so written guidance should be considered as just one tool that can be used as part of a collective of measures e.g. formal training, mentoring etc

Examiner attendance at standardisation meetings

All of the participants of the study were conscious of the need for examiners to gain a common understanding of mark schemes and in this regard standardisation meetings can be very effective. The standardisation meetings also allow the mark schemes associated with a particular examination sitting to be scrutinised much closer in terms of content and it is where the examiner's knowledge of the subject becomes an important part of the process as they have the opportunity to remove, add or amend the mark schemes before candidates / students are affected. Additionally, if needed the content of the mark schemes can be altered significantly or minor amendments made dependant on as to how the candidates / students have interpreted a question.

Although both AB1 and AB2 were insistent on an examiner attending a standardisation meeting prior to marking examination scripts they differed greatly in how much active participation they required examiners to have during the course of the meeting. Indeed Alan was somewhat dismissive in his views as to how much reliance or need they placed on examiners input to mark schemes. Jane however welcomed input, within reason, to encourage and enforce that common understanding of mark schemes and ultimately to aid reliability.

When researching what examiners felt of standardisation meetings Baird, Greatorex, and Bell (2004) found that examiners considered all aspects of the standardisation process to be important, particularly gaining the knowledge as to how a principal examiner wished a mark scheme to be interpreted. As discussed in the literature review Thyne (1974) identified that if mark schemes are precise and clearly state where marks can and should be awarded, then the more relevant and consistent marking is likely to be.

The research would suggest however that attendance at standardisation meetings does not necessarily improve the marking of examiners and that perhaps the approach taken by PB3 e.g. remote standardisation, would suffice. Although it is important to remember the other benefits attendance at standardisation meetings can bring. In addition to gaining a valuable understanding of the mark schemes, the research demonstrates that examiners attending standardisation meetings gained confidence in their marking and in the knowledge that they were marking appropriately and that they had the same understanding of the mark schemes as that of their fellow examiners and principal examiner.

Whether standardisation is done remotely or via face to face meetings the purpose of standardisation is essential in ensuring fair and transparent assessment, that is all examiners approaching mark schemes with the same common understanding and awarding marks where agreed and warranted.

Standardisation meetings effectiveness in aiding reliability

It was acknowledged by Alan and Jane that one of the most important factors in standardisation meetings aiding examiner reliability was good chairmanship by a 'principal examiner'. The role the principal examiner plays is to ensure (as far as practicable) that examiners are in agreement on what marks should be awarded for e.g. case law, must prove a legal point to be awarded a mark rather than just the case name just being supplied which in turn would not warrant a mark. They also play an important role in facilitating agreement on marks to be added / removed from a mark scheme and ultimately if required making a final decision to overrule overbearing and dogmatic examiners. In essence, giving high level direction but also allowing free flowing discussion and critique of questions, mark schemes and the responses provided by students. Both Alan and Jane felt very strongly that without this firm chairmanship, standardisation meetings would not be affective in aiding reliability and the meetings would be controlled by the more verbose members of the marking team. Indeed Alan felt that even very good markers would "come off the rails" if standardisation meetings did not occur.

Interestingly although Alan and Jane both felt that meetings had to be controlled, group dynamics do play a part in gaining a common agreement on a mark scheme particularly when two examiners are in disagreement and it is up to the Chair to manage that disagreement in a controlled way. As Jane stated, standardisation is

> "…not about personal opinions as such, it's literally the standardisation, and the agreement that everybody in the room knows how to apply that mark scheme".

It would appear therefore that without the correct selection and appointment of a principal examiner, standardisation meetings could be less affective. The literature e.g Desforges (1989), QCA (2008) major on the professional qualifications and experience of marking that a principal examiner must hold but do not make reference to the personal characteristics that would make the role holder most effective. Undoubtedly a principal examiner needs to be credible in terms of knowledge and understanding of the subject being examined but it would appear that they also need to be empowering, perceptive and a good influencer.

None of the participants involved in this study made reference to the observations made by Baird, Greatorex, and Bell (2004) that suggested that marking reliability is purported to be produced by having an effective community of practice, that is a network of people who have a shared 'project' (activity) which they continually renegotiate. Undoubtedly, standardisation meetings are a rewarding enterprise for examiners in that they meet with their peers and are challenged intellectually. It was not proven in this study however that a 'community of practice' positively affects an examiners marking positively or that it makes the marking process more rewarding.

Something which was not identified during a review of the literature (perhaps because of its fundamental nature) but identified by Jane, was that the facilities provided for the standardisation meeting need to be conducive to holding a meeting in terms of comfort (both seating and temperature), space and lighting and this makes a great deal of sense.

Post standardisation monitoring of examiners

Another aspect of the study emphasised the value placed on post standardisation monitoring. It is fairly common practice (Greatorex and Bell 2008), although a relatively new innovation, to require examiners to supply a sample of their marking following standardisation.

AB1 and AB2 both adopt a team leader approach once marking has commenced which involves examiners returning a sample of their marked scripts to be over-marked by their respective team leader or if numbers allow a principal examiner. The examiners then receive feedback on their performance, usually by telephone and then email. The examiners also receive their scripts back so they can tangibly see were they gave marks compared to their team leader or vice versa. It was felt that in the first instance any critiquing of an examiners marking should be done in a very personable style to encourage rather than berate. If however marking was considered to be significantly outside of acceptable thresholds another sample would be requested.

For reasons of reliability and validity both AB1 and AB2 encourage their examiners to postpone further marking of scripts until feedback is given to them on their marking. This is requested as they may need to make an adjustment to how they interpret the making process. This postponement in marking makes it essential that feedback from team leaders is prompt so as not to place additional time pressures on examiners particularly in respect of the tight marking windows which in turn could have a detrimental affect on their marking as discussed previously.

Although both awarding bodies attempt to support examiners as much as possible and take reasonable and expensive e.g. group meetings (fuel costs, attendance pay, overnight accommodation etc), steps to ensure examiners should have a common agreement and understanding of mark schemes, those examiners who fail to demonstrate this through their marking sample are placed under no illusions that they would have their scripts recalled and retraining offered. To permit an examiner to continue to mark inappropriately is not only unfair to candidates but it also has other implications such as distorting examination pass rates and to cause additional marking post moderation, under very tight deadlines.

If the marking is sufficiently reliable the senior examiner deems that the examiner can continue to mark as before. If the marking is not sufficiently reliable then:

> ”…examiners are required to provide a further sample for review and receive more feedback, and sometimes stronger action is taken depending upon he circumstances” (Greatorex and Bell, 2008, p. 334).

Moderation

It could be argued that no examination is perfectly valid, particularly in respect of marking-consistency and sampling of questions, and that small divergences from the standard can be expected. Moderation is therefore an essential part of the examination process in reviewing and ratifying the marks given by examiners. As Torrance (1996) alluded, moderation understood as review is essential in monitoring assessments and to ensure that they are fair, to check that procedures are adhered to, and to confirm on interpretations – that is, how mark schemes have been applied.

It is the role of a moderator is to ensure that that examination results released to candidates are, as far as possible, fair to all concerned and that the standard of the marking in terms of the degree of discrimination applied by examiners is; in the belief of the moderator, appropriate.

Both AB1 and AB2 undertake moderation post marking and consider it to be a key part of the process. Although examiners marking would have been reviewed already by team leaders, it was felt that moderation gives the opportunity to gain a better overall view of how the examiners had performed and as Alan suggested to inform "…how much intolerance is still in the system".

Additionally moderation gives Standards Officers and principal examiners the opportunity to see marking performance in more detail and to gain a closer association with an examiners abilities and traits e.g. harsh and over lenient marking. It is useful to have this familiarity with an examiners marking as a decision can then be made as to whether their scripts will need closer scrutiny or if efforts can be concentrated elsewhere.

When moderating the examiners scripts AB1and AB2 take a look at a range of marks i.e. pass, credit, distinction to decide whether a mark adjustment should be applied e.g. plus 2 marks to each scripts an examiner has marked if they have been found to be marking slightly harshly. AB2 gave a tolerance of 5% on were a team leader would have marked a script compared to an examiner, if marking is found to have fallen outside of this tolerance a whole batch is remarked and training offered to the examiner before being allowed to mark again. AB1 took a slightly harder line with the approach that should an examiner be found to be marking inappropriately a contract would not be subsequently offered to that examiner for marking any future examinations. PB3 are also prepared to make mark adjustments *if* an examiner is found to be marking aberrantly during the standardisation process and their scripts are reviewed as a result. As a rule however AB3 do not undertake moderation in any significant sense.

Alan observed that on occasion moderation may identify that marking may be slightly at variance due to a number of factors other than an examiners ability to mark accurately. For example it may be that a question was particularly difficult to mark or that because the way a question was worded candidates have approached it differently. If the question setting process and standardisation of mark schemes are undertaken appropriately, Alan’s observations above should be of relatively minor significance in relation to examiner reliability. That is questions should be well designed and if there issues identified with them then it should be addressed at standardisation and mark schemes adjusted to reflect this.

Strengths of the Study

A strength of this study is the qualitative style of research adopted for the study as it produced some very personal and in depth data from the participants on a large range of issues in connection to examiner reliability in a very practical sense. Although this was a small scale study the data was rich in content and experience.

I feel that I was able to isolate my own experiences of examining and disengage with my own views and experiences and focus on the interviewee’s perspective in an effort to prevent bias in the interaction.

The participants all entered freely in to the study and were well informed of the basis and intention of the research. It was made clear to the participants what their involvement would mean leading to informed consent being given.

Having three influential and experienced ‘managers’ of vocational examiners, it gave a range of perspectives on the subject and a varying view as to what affects their reliability.

Limitations of the Study

The research was limited by being a small scale study, had there been resources to interview more subjects and possibly to use surveys in conjunction with the qualitative approach this may have given more conclusive larger scale results. The participants could have been from a wider variety of approaches, for example examination boards from outside of the health, safety and environmental management field, to give the data a more diverse and comprehensive view.

It would also have been beneficial to have interviewed examiners from different genders and marking cultures to see what they believe affects their reliability when marking so as to have made the data even more diverse.

As I used semi structured interviews for data gathering I could have biased the answers of the respondents with my tone of voice, my body language and the wording of the questions.

Conclusion

Conclusion

The aim of this study was to discover what affects the reliability of ‘vocational’ examiners involved in the marking of public examinations. This was prompted by an awareness that there were variances in how examiners interpreted and applied mark schemes and therefore how they marked candidates work. I believe the aim has been achieved. The themes that were derived from the data supplied by the sample were diverse and coupled with the literature reviewed, interesting, and it has given me a depth of understanding that has ignited my enthusiasm and a need to discover more.

Examiners play a major role in ensuring that the qualifications that are awarded to those sitting public examinations are just and fair. Those examinations that have not been marked appropriately, can have a detrimental effect on the future prosperity of candidates if their examinations are under-marked, and give an unfair advantage, to those who do not deserve it, if over-marked. In addition to disadvantaging or benefitting those sitting an examination, aberrant marking can also affect the integrity of an award and / or qualification by inflating or deflating pass rates which can in turn result in the perception of the value of a qualification being called into question. In relation to those candidates undertaking the awards involving the participants of this study, there can be potentially disastrous results in terms of the health and safety of both workers and the general public and also possible damage to the environment. The reason for this is that when qualifications have been awarded, the generally held perception is that knowledge and application at a suitable academic level has been achieved as has a defined level of professional competence. This certainly is demonstrable in relation to specialist vocational awards i.e. those offered by the sample. Those attaining the higher level qualifications offered by the sample are seen as having a certain level of competence, indeed on the attainment of the Level 6 health and safety awards offered by AB1 and AB2, application towards gaining 'chartered status' as a health and safety practitioner can be progressed through the Institute of Occupational Safety and Health. It must be noted however that the organisations used as part of this study would not suggest that academic achievement with them necessarily assumes competence.

What was of particular interest within the study was that the majority of the literature researched, with the exception of the Qualifications and Curriculums Authority (QCA) (2008) GCSE, GCE and AEA Code of Practice, majored on the theoretical issues surrounding the use of examiners such as the cognitive marking strategies that examiners adopt when examining, the levels of educational achievement versus practical experience of marking and the resultant reliability and how a positive community of practice can aid the validity and reliability of marking. 'Alan', 'Jane' and 'Sally' however were much more concerned with the practicalities of managing examiners reliability and the very real challenges that examining on a large scale presents although indirectly they were in agreement with the literature reviewed.

The two awarding bodies, AB1 and AB2, approached the use of examiners in much the same way which was not surprising as they both offer qualifications accredited by the Qualifications and Curriculum Authority; however Alan and Jane's view of the usefulness of the examiners in the setting of standards and mark schemes was very different. Jane encouraged and welcomed debate during standardisation meetings to engender the 'common understanding' of examiners and the mark schemes they were to use. Alan preferred to limit the active participation of examiners to predetermined marks that he was already prepared to change in advance of the meeting to keep them "motivated". Indeed Alan's view of examiners was at times belittling of them ("little examiners") and their importance in the examination process.

PB3, although open to scrutiny in how they manage some aspects of the examination process, do some things very competently, for example the pass boarder review that they undertake is significantly wider than that undertaken by the two awarding bodies. That is they ensure, as far as practicable, that everybody deserving of a pass, should achieve one and to a certain extent those that do not deserve to pass, don't. In light of the study undertaken by Baird, Greatorex, and Bell (2004) the fact that they do not hold standardisation meetings per se does not necessarily mean that their examiners have any less of a common understanding of mark schemes than the examiners used by the two awarding bodies who go to great expense to ensure and stipulate attendance at the meeting as a prerequisite to examining.

All three of the sample demonstrated that their organisations considered the examination process to be of huge importance in maintaining the credibility of the qualifications / awards that they offer and to this end it would be advisable to keep a watching brief on the new technologies which are becoming available e.g. online marking. The advantages of being able to monitor examiners in 'real time' would be considerable in gaining a truer perspective on examiners marking well before candidates are affected.

What was of immense interest was that the study showed that 'vocational examiners' do not need to be managed any differently from an experienced educationalist who may be examining, when marking public examinations involving high numbers of scripts. All examiners are fallible and some, based on the responses provided by the sample, do not have an aptitude for examining or are unable to follow instructions.

Based on the findings of the study in relation to both the literature researched and the views of the participants involved in the sample, the following recommendations can be drawn as aiding examiner reliability and thus ensuring a robust and valid examination process:

- There appears to be no correct cognitive marking strategy that should be stipulated for use by examiners when marking a particular question, as there is no correspondence between the method of marking used and any resultant accuracy;
- ‘Horizontal marking’ should be encouraged when marking large numbers of scripts as by marking one question at a time from the whole batch (horizontally) it was felt that examiner’s have less opportunity to become aggrieved by a candidate’s handwriting or approach to an examination paper;
- Although inter-rater reliability is obviously desirable amongst examiners e.g. consistency amongst them. It is essential that any examiners used are marking consistently (not necessarily with perfect accuracy) so scripts can be appropriately adjusted post moderation;

- Examiners can and should be informed to self regulate. That is not to mark scripts if fatigued as marking performance will deteriorate (affecting intra-rater reliability) as was demonstrated by the DRIFT phenomenon;
- Examination scripts should be randomised before issue to examiners and should also be anonymised e.g. identified by a candidate number as examiners, as identified by Aslett (2006), sometimes display emotional factors which can play a part in affecting reliability. That is, where an examiner is aware of a student's identity, their marking can potentially be profoundly affected;
- Suto & Nádas (2008) generalised that marking could be affected by both the demands of the marking task, including marking strategy complexity, and a marker's personal expertise. It can therefore be argued that accuracy of marking can be improved both by reducing the demands of the marking task and by increasing a marker's personal expertise in the marking process and knowledge of the subject under examination;
- The research demonstrated that it is advisable to use examiners with a high level of academic achievement, in the case of the sample, the qualifications should be in the field of in the health, safety and environmental management, as is it essentially a

better predictor of accuracy of marking than either teaching or marking experience;

- To reduce the scope for variability amongst examiners, marking should be undertaken by the minimum possible number of examiners taking into account the nature of the unit/component being assessed and the time required to mark candidates' work. Any time given for the marking of scripts should be realistic;
- Training of examiners before they examine for the first time should be undertaken as it will aid in bringing examiners' differences in leniency (inter-rater reliability) to an acceptable level once they examine 'live' papers;
- Examiners should be issued with printed guidance / reference procedures stipulating the 'rules' of marking clearly stating how to and when to award marks;
- Post standardisation but pre moderation of examiners by team leaders should be undertaken as a quality check that the correct interpretation of the mark schemes has been understood and applied by examiners before large numbers of scripts have been marked. Those that have been marked can be revisited by examiners as a result and adjusted;
- The examiners should be guided through the examination process by the provision of appropriate leadership e.g. a principal examiner. The role of the principal, in terms of

affecting examiner’s reliability, is to give clear guidance as to how mark schemes should be interpreted and when and where marks should be allocated;

- Whenever possible mark schemes should involve the Unit-Counting method of marking rather than levels-of-response marking as it is most likely to produce the most consistent marking amongst examiners. Mark schemes should also be clear and easily interpreted;
- Attendance at standardisation meetings does not necessarily have to be a prerequisite of examining for a particular sitting in terms of aiding reliability but examiners have expressed that they find it rewarding, according to the research undertaken by Baird, Greatorex, & Bell (2004) in terms of helping them to understand and interpret mark schemes; and
- Moderation is a key part of the examination process in aiding reliability, with the principle being that of safety in numbers e.g. the more people who agree about the overall grade awarded to a particular performance the more valid the assessment is likely to be.

Possible Future Research

Other themes emerged from the research which do require further exploration to gain a better insight into the role of examiners, but which were not able to be explored in this dissertation are:

- The drivers in becoming an examiner;
- The recruitment and retention of examiner’s;
- Feedback to examiner’s and its affect on reliability;
- An examiner’s role in Identifying plagiarism; and
- The importance of clerical checks on examiner’s marking.

APPENDIX 1

References

Baddeley, A. (1999). Essentials of human memory. Hove:Psychology Press.

Baird, J., Greatorex, J., Bell, J.F. (2004). What makes marking reliable? Experiments with UK examinations. Assessment in Education Principles, Policy and Practice 11(3), 331–48.

Bell, J., Bramley, T., Claessen, M., Raikes, N. (2007). Quality control of examination marking. Research Matters:Cambridge Assessment Issue 4, 18-22.

Bell, J. (2005). Doing your Research Project (4th ed.). Maidenhead: Open University Press.

Bogdan, R., Taylor, S. (1975). Introduction to Qualitative Research Methods – A Phenomenological Approach to Social Sciences. New York: John Wiley & Sons.

Bond, T., (2004). Ethical Guidelines for Researching and Psychotherapy. British Association for Counselling and Psychotherapy.

Denscombe, M. (1998). The Good Research Guide. Buckingham:Open University Press.

Desforges, C. (1989). Testing and Assessment. London:Cassell Education Limited.

Ecclestone, K. (2001). “I know a 2:1 when I see it”: understanding degree standards in programmes franchised to colleges. Journal of Further and Higher Education, 25, 301-313.

Filer, A., Pollard, A. (2000). The Social World of Pupil Assessment – Processes and Contexts of Primary Schooling. London:Conitnuum)

Gipps, C. V. (1994). Beyond Testing Towards a Theory of Educational Assessment. London:The Falmer Press.

Greatorex, J. (2007). Did examiners’ marking strategies change as they marked more scripts? Research Matters:Cambridge Assessment Issue 4, 6-13.

Greatorex, J., Bell, J. (2008). What makes AS marking reliable? An experiment with some stages from the standardisation process. Research Papers in Education 23 (3), 333–355.

Greatorex, J.,Bell, J. F. (2004). Does the gender of examiners influence their marking? Research in Education:University of Cambridge Local Examinations Syndicate. 1-18.

Greatorex, J., J., Baird, Bell, J. F. (2002). ‘Tools for the trade’: What makes GCSE marking reliable? Learning Communities and Assessment Cultures: Connecting Research with Practice.

Hamp-Lyons, L. (1990). Second Language Writing: Assessment issues. Research Insights for the Classroom Norwood:Ablex Publishing Corporation. 127-153

Klein, J. (2003). Impairment of teacher efficiency during extended sessions of test correction, European Journal of Teacher Education, 26 (3) 379-392.

Kramer, A., Coles, M., Logan, G. (1996). Converging operations in the study of visual selective attention. Washington DC: American Psychological Association.

Kvale, S. (1996). Interviews; An Introduction to Qualitative Research Interviewing. London: Sage.

Laming, D. (2004). Human judgement: the eye of the beholder. London:Thompson.

Lewis, D. (1974). Assessment in Education. London:University of London Press.

Mason, J. (2002). Qualitative Research (2nd ed). London: Sage.

McLeod, J. (1994). Doing Counselling Research. London: Sage.

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

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

Newton, P. (1996). The reliability of marking General Certificate of Secondary Education scripts: mathematics and English, British Educational Research Journal. 22, 404-420.

Office of the Qualifications and Examinations Regulator (Ofqual). Retrieved January 01, 2009, from http://www.ofqual.gov.uk/191.aspx.

O'Neill, G. (1985). Self, teacher and faculty assessments of student teaching performance: a second scenario. The Alberta Journal of Educational Research. 31(2) 88-98.

PB3's website, accessed 2nd May 2009

Pidgeon, D., Yates, A. (1968). An introduction to educational measurement, London: Routledge and Keegan Paul

Pinot de Moira, A., Massey, C., Baird, J., Morrissy, M. (2002). Marking consistency over time. Research in Education. 67, 79-87.

Qualifications and Curriculum Authority (2008) Review of question paper setting and senior examiner training for GCSE and A level. London:QCA 1-44

Qualifications and Curriculum Authority (2008) (QCA) GCSE, GCE and AEA Code of Practice Qualifications and Curriculum Authority: London

Ruth, L., Murphy, S. (1988). Designing writing tasks for the assessment of writing, Norwood NJ:Ablex

Spear, M. (1997). The influence of contrast effects upon teachers' marks. Educational Research. 39, 229-233

Suto, I., Nádas, R. (2008). Towards a new model of marking accuracy - An investigation of IGCSE biology, , Research Division:Cambridge Assessment. 1-13.

Suto, I., Nádas, R. (2008). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers Research Papers in Education. 23(4) 477-497

Times Online (2009) Computers to mark English tests. Retrieved September 25th 2009 from http://www.timesonline.co.uk/tol/life_and_style/education/article6848572.ece

Thyne, J. M. (1974). Principles of Examining. London:University of London Press

Torrance, H. (1996). Evaluating Authentic Assessment. London:Open University Press.

Valentine, C. M. (1932). The reliability of examinations: an enquiry with special reference to the entrance examinations to secondary schools, the school certificate examinations, and the award of scholarships at universities. London:University of London Press.

Weigle, S. (1998). Using FACETS to model rater training effects. Language Testing. 15, 263-287.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. Language Testing. 10(3), 305 335.

Wolf, A. (1995). Competence based assessment Buckingham:Open University Press.

Wolfe, E.W., Moulder, B.C., Myford, C. (2001). Detecting Differential Rater Functioning over Time (DRIFT) Using the Rasch Multi-faceted Rating Scale Model. Journal of Applied Measurement 2(3). 256-280.

APPENDIX 2

MA in Education (by research) Dissertation Consent Form.

In order to complete a Masters Degree in Education (by research) at the University of York, I am undertaking a qualitative research project examining what affects the reliability of vocational Examiners.

I, the undersigned have read and understood the dissertation proposal and explanatory letter, and agree to participate in an interview with Matthew Powell-Howard. I consent to the following terms;

- The research is looking at what effects the reliability of vocational examiners.
- The interview in which I partake will be recorded and when deemed necessary as part of the dissertation, transcribed.
- My anonymity will be ensured at all times.
- Confidentiality will be maintained.
- Quotations in the research project will be anonymous and a pseudonym will be used when required.
- I am able to withdraw from the interview at any time.
- I am aware that the information given will be analysed for academic research.
- The research is carried out within the ethical framework for research as specified by British Educational Research Association document *'Revised Ethical Guidelines for Educational Research'*.

Name……………………………………………………………………………

Signed…………………………………………………………………………..

Date……………………………………………………………………………

APPENDIX 3

Dissertation Proposal

Area of Enquiry

My area of enquiry is related to what affects the reliability of vocational examiners. By vocational I mean subject experts rather than educationalists. I am interested in this area, as during my time as both an Examiner and as a Standards Officer for an awarding body specialising in vocational awards, I was aware that there were variances in how Examiners approached mark schemes and therefore how they marked candidates work. Examiners' marking to a common standard and a common interpretation of mark schemes is important so as to not disadvantage or favour clusters of students which could affect the integrity of an award and / or qualification. Based on personal experience, I am also aware of the lack of training and monitoring Examiners undergo; I was considered to be a suitable examiner based purely on my professional qualifications, experience and Chartered status within my field of expertise and my lack of experience in relation to marking and educational practices was not considered.

Although there has been extensive research into what affects marking e.g. increased monitoring, clearly structured mark schemes, little research has been undertaken as to what awarding bodies using subject experts have found to be most effective in improving the reliability of their Examiners. There may also be differing work practices relating to Examiners amongst different awarding bodies.

Sources

My secondary sources will be journal articles and training literature. The areas that I am researching associated with Examiners would include personal characteristics e.g. gender, age and experience in addition to training, mentoring, monitoring and published guidance from educational bodies i.e. OFQUAL (Office of the Qualifications and Examinations Regulator) in relation to reliable assessment. Supporting research questions will be how training and mentoring contributes to the reliability of Examiners.   Vocational Examiners specifically are not a group whose experiences of marking have been well researched. However, research on factors influencing reliability of marking exists and this will be useful in contextualising the present study, as well as indicating previously-identified examples of best practice in marking.

Methodology and Design

I intend to use qualitative methods of research in order to gather data focusing on current practices and the experiences of Examiners. Using semi structured interviews I believe I will acquire an understanding of how different Awarding Bodies and professional bodies offering qualifications measure the reliability of their Examiners. The people I have identified as being experienced enough to tell me how Awarding Bodies / professional bodies manage their Examiners are either the Principal Examiners or the Standards Managers of the individual Awarding Bodies that I have identified as a representative sample. I have decided to define my sample of organisations as three of the largest and most respected providers of health safety and environmental management qualifications in the UK, and who will generally rely on health, safety and environmental professionals to examine on their behalf. The sample thus consists of:

[redacted]

Following this initial consultation I may then need to speak to Examiners who are both new to the role and well established to ascertain what they feel affects their reliability / performance. (example references Kvale, 1996; Bell, 2005)

Data Collection

Semi structured Interviews will be audio recorded and if / when required transcribed.

Literature review

As previously alluded to there has been a large amount of research dedicated to examining and ensuring the reliability of examiners although some, but not all, make the assumption that those who are marking are lecturers, teachers, graduates etc rather than Examiners who are not working in the field of education. *Example* materials are:

- Issue 4 Research Matters, Cambridge Assessment Agency 2007
- Does the gender of examiners influence their marking? (Greatorex / Bell 2004)

- What makes marking reliable? Experiments with UK examinations. Baird, J-A., Greatorex, J. and Bell, J.F. (2004)
- GCSE, GCE, VCE, GNVQ and AEA Code of Practice 2007/8

Data Analysis

This will be drawing the threads and themes of the data (Silverman, 2000) to gain some understanding as to what affects the reliability of Examiners.

Ethical and Professional Issues

The ethical issues are the potential harm that can be caused to an examination board should information on how they manage their Examiners be exposed to candidates, particularly if the organisation is not seen to be following best practice. As I am a direct employee of one of the organisations I would need to keep my boundaries as a researcher, and be very open about my employment to the two other awarding bodies. I must also respect the confidentiality and commercial sensitivity of any documentation, or other information, that I may be privy to.

Confidentiality for the participants is a consideration and also acquiring informed consent is necessary, ensuring that all the interviewees understood how the data was to be used and that they would be informed of any use other than that of the dissertation e.g. publication.

Anticipated outcomes

I believe that I am objective on this topic of research, however I would expect that those vocational examiners who have undergone training, attend standardisation meetings to agree clearly defined mark schemes and have the support of mentors should mark reasonably accurately. I would also expect that awarding bodies follow published guidance in relation to the recruitment and monitoring of Examiners more than they once did.

The results of my study could be used for giving guidance to those organisations that use ‘subject experts’ as Examiners in developing practices that would improve the effectiveness and reliability of them.

Results

The results will be seen by my tutors. I am not planning on producing an anonymised report to participating organisations although if requested I would do so.

Supervision

My academic supervisor is Dr Vanita Sundaram at the University of York.

APPENDIX 4

Sample of transcripts

‘Sally’ 29th April 2009

Interviewer: Erm, so is there any form of interview process when you are selecting Examiners?

Respondent: There isn’t, although its something we’re looking at. Erm, we don’t do a formal interview, after the scanning of the CV and then meeting the standard requirements, if they pass that stage they would then be invited to either the Level 3 training workshop or the Level 6 training workshop. But it is something we are in discussion with HR at the moment to do more in this with regards to thorough selection process.

Interviewer: Right, so if somebody becomes an Examiner, erm, how many scripts do you tend to give them?

Respondent: That varies dependant on the qualification that they are marking, for example, for this typical qualifications, it would be 72 to 80.

Interviewer: And do you think, its proportionate their reliability as an examiner based on the more scripts they get, the more consistent they are or do you think, do you give them that sort of number, purely through necessity, or is it a ……..

Respondent: The balance is the time that they’ve got available to mark, and from a reliability point of view, obviously the fewer examiners you have, the better. But set against that, we’ve got to look at how many scripts there are to mark and the number and available marking days the Examiners have. So it’s a balance between that and keeping the number of Examiners as low as possible.

Interviewer: Okay. What do you feel is a barrier and also conversely a driver to reliability in terms of deadlines, remuneration, erm, the time they've got to mark? Do you think that sometimes can affect someone's reliability or do you think it can drive someone's reliability?

Respondent: I think both, I think depending on the personality of the individual. Certainly the timescales can potentially affect reliability. So if an examiner isn't allocating time across there, what we try and give is a three weekends. If they're leaving that all to the end, and the rushing and obviously that is going to affect their reliability. Erm, but the converse of that, I can remember when I first started marking, some examiners will go over the same scripts again and again, because they are so worried about not applying the mark scheme fairly, but that can almost conversely affect reliability. What were the other two you mentioned?

Interviewer: A driver, I mean they were really to act as prompts really. But what do you think is a driver towards marking reliably? Erm, I was thinking of, erm, team leaders, you know somebody else is going to take a review of the script, perhaps that makes you be a little bit more conscientious. Again, that's a personal view, but I wondered if you have any, or, I mean, sometimes, it could be "gosh I'm in a really responsible position here, I will take some time over this".

Respondent: I think it's that, and that's the quality we would be looking for in an examiner regardless of the remuneration and the benefits they get from being involved in the process. The ultimate driver should be, the fairness of the assessment to the candidate, and I think for most of the examiners we have that, that does apply. Occasionally, we get comments or feedback that examiners are not paid enough to give due time and resource to the marking process, but with a team leader process being introduced, we're hoping that we're weeding out the examiners that are not giving it the due time and attention.

Interviewer: Okay. So what training, if any, do you give potential new examiners? And, what's the objective of this training?

Respondent: That is a mandatory requirement to them to actually mark live scripts. You are required to come to a one day training workshop, where they get a background to NEBOSH, a background to assessment procedures and then most of the time is devoted to almost replicating a standardisation meeting, so they understand that process. And some actual marking of scripts. So they start to understand how to mark and annotate scripts appropriately.

Interviewer: Okay.

Respondent: But also it gives the attendant Standard Officer a feel for their ability to mark but also the potential attitude, and as you were saying earlier, the confidence and the application of how they perhaps can perform in future standardisation meetings.

APPENDIX 5

Interview schedule version 2

- How do you attract subject experts to the role of an “Examiner”?
- What do you feel attracts individuals to the role?
- What do you look for in any potential Examiners:

a) Professional qualifications;

b) Personal characteristics?

- Is their any form of interview process?
- Approximately how many scripts do you give Examiners?
- Is this because you feel more / less scripts aids reliability / accuracy?
- What do you feel is a barrier and also conversely a driver to reliability (deadlines, remuneration)?
- What training, if any, do you give to potential new Examiners and what is the objective of this?
- How would you deem the training to have been successful?
- Do you have any marking guidance or literature that you give to Examiners to refer to?
- What meetings are Examiners expected to attend e.g. standardisation meetings?
- In standardisation meetings how do you assess Examiners interpretation of mark schemes?
- What do you feel aids an effective standardisation meeting (environment, chairmanship, group dynamics)?

- What type of behaviour within the standardisation meeting enables you to feel confident of Examiners understanding of mark schemes?
- How do you monitor Examiners once the marking process has commenced?
- If you use Team Leaders / mentors how are they selected?
- How are the Team Leaders / mentors monitored for reliability?
- Post marking, what checks on the Examining is undertaken?
- Do you provide feedback to Examiners on their marking performance?
- Are there any improvements to your current system that you would like to develop?
- Do you follow any published guidance in relation to monitoring Examiners such as the GCSE, GCE and AEA Code of Practice?
- On the successful completion of your award(s), what are the benefits to candidates in terms of job prospects, salaries etc?
- I’m coming to the end of the interview, is there anything you feel you would like to add?