PUBLIC PERCEPTIONS OF RELIABILITY

Suzanne Chamberlain

OFQUAL/10/4708 **January 2010**

This report has been commissioned by the Office of Qualifications and Examinations Regulation.

CONTENTS

Tables

1 EXECUTIVE SUMMARY

This report forms part of Ofqual's Reliability of Results two-year research programme (Strand 3: 'Public Perceptions of Reliability'). The aim was to conduct a number of focus groups to gauge public awareness of assessment reliability, and the degree to which the public wishes to be (better) informed of reliability issues. Participants were also asked to comment upon the means by which such issues may be best communicated. Ten focus groups were conducted with between five and ten participants in each group (74 participants in total; 28 male and 46 female). Participants were grouped by occupation or economic status with two groups each of job-seekers, employees, employers and PGCE students, one group of primary school teachers and one group of secondary teachers from a selective school.

The main findings can be summarised as follows:

- With the exception of the secondary school teachers, the participants had limited awareness of the concept of reliability. Participants were able to recognise forms of human error in the assessment process but often failed to envisage how this might impact on the reliability of their assessment outcomes.
- The participants struggled to see how measurement inaccuracy (Newton, 2005) could be termed 'assessment error' and how it could impact on the reliability of outcomes. Instead, they suggested that measurement inaccuracy was an inevitable part of life, and that to draw attention to its impact on assessment outcomes would not be beneficial.
- The participants had rarely questioned the reliability of the assessment process or their assessment outcomes, and showed a significant amount of trust in the system to award them the 'right' outcomes. Some participants had experiences of re-marks or appeals. This appeared to make them more questioning of the accuracy of their results than other participants, but seemed to do little to undermine their trust in the assessment system as a whole. The secondary school teachers spoke extensively about their experiences of challenging students' results, and demonstrated their awareness of how errors could occur.
- Participants tended to trust examiners to assess their work fairly, believing that examiners are professional and well trained subject experts. The participants could recognise, however, that some subjects require more interpretation than others, and thus that the reliability of marking could be variable. The secondary school teachers tended to be less trusting – many acted as moderators themselves in order to mediate the influence of external examiners, and to gain a better understanding of assessment criteria to pass on to their students.
- On the whole, the participants suggested that they would like to be more informed about assessment reliability, but only through a better understanding of how the assessment process works i.e. knowing what happens to a candidate's script after the candidate has completed the examination. There was a notable lack of support for any quantification of reliability and, in particular, publishing a reliability statistic alongside a candidate's grade. The secondary school teachers were particularly emphatic that any initiative to enhance understanding of reliability should begin with teachers and students, and not with parents or the public at large.

As the literature suggested would be the case, the participants found assessment reliability, and in particular measurement inaccuracy, difficult concepts to comprehend. Active discussion about reliability was achieved through the sharing of ideas and experiences and the use of vignettes (short case studies and scenarios). In the light of this, and the participants' comments, it is suggested that a qualitative approach to reliability that focuses on

students and teachers may be a possible way forward in enhancing the dissemination of reliability information. It is acknowledged, however, that this finding may be a product of the questions asked or the method by which they were asked. The Strand 3 questionnaire survey may uncover greater support for a quantitative approach to the dissemination of reliability information.

2 INTRODUCTION

The outcomes of educational assessment serve several high-stakes social functions. They have become increasingly important indicators of the performance of teachers and educational centres at a local, national and even international level, with all that that entails for student and staff recruitment, retention and funding (e.g. Goldstein, 2001). For the candidate, assessment outcomes act as 'entry tickets' to the labour market or education and training opportunities, and as such have a significant impact on an individual's life chances (Denscombe, 2000). These weighty social functions imply that assessment outcomes must have a great deal of dependability and credibility. Newton (2005) suggests, however, that reliance on an assessment grade as an accurate measure of ability represents a myth of perfect assessment and perfect reliability. This goes hand-in-hand with a belief that errors, although they may occur, are not intrinsic to the measurement process. Instead, it suggests that someone (an examiner or question paper writer) or some agency (an awarding body) is culpable when an assessment outcome is questionable. This view fails to acknowledge the fact that grades are an approximation of ability – rather than 'truths' in themselves - and should be interpreted within the context of the reliability of the measurement process (Murphy, 2004).

Several commentators have argued the case for reporting some measure of reliability alongside assessment outcomes (e.g. Newton, 2003; Wiliam, 2003). This could take the form of confidence intervals (Newton, 2003), a range of grades within which a candidate could confidently be placed (Please, 1971) or reliability coefficients and standard errors of measurement (American Educational Research Association, American Psychological Association and the National Council on Measurement in Education, 1999). Arguably, reporting reliability statistics would allow qualification-users to determine for themselves whether an outcome was sufficiently accurate for an intended purpose.

Research has shown, however, that stakeholders and the general public do not fully understand technical assessment concepts such as reliability (Newton, 2005; Taylor, 2007). Moreover, a recent study conducted as part of Ofqual's Reliability of Results research programme suggests that members of the public have only a passing interest in understanding technical issues, and only so far as it is in direct relation to their own lives (Ipsos MORI, 2009). There is a tension then, between (a) discussing the reliability of national assessment outcomes while acknowledging that some measurement error is inevitable, and (b) doing so in a way that is engaging and accessible to the general public and likely to foster a more informed interpretation of results.

Public opinion tends to follow a developmental pattern that begins with the emergence of a topic, followed by increased public attention and discussion, which then results in opinion change or the disappearance of the topic from the public domain (e.g. Glasser and Salmon, 1995). The issue of assessment reliability is only sporadically on the public agenda (and in non-specific terms), mostly triggered by the summer publication of results and the accompanying media attention. The focus tends to be on avoidable errors or failures in the

system (e.g. the 2008 SATs marking crisis) or debates about standards. The sporadic nature of this attention and the focus on negative events means that public understanding of reliability issues and the inevitability of measurement error is rarely progressed – instead the issue disappears from the public domain until the following year or the next crisis.

The Ofqual Reliability of Results research programme seeks to address, among other things, the question of whether publishing information about assessment reliability would be beneficial to the educational community and the public at large. Newton (2005) argues that a lack of understanding about assessment reliability underpins and fuels the negative media attention, which in turn erodes public confidence in assessment systems. To overcome this, various qualification-user groups require more information in order to better interpret and apply assessment outcomes. Newton (2005) suggests that this will increase, rather than undermine, public confidence in the assessment process as the public becomes more aware of what assessment systems can be reasonably expected to deliver, and what they are unable to deliver (e.g. outcomes that are free of measurement error).

This study seeks to explore the degree to which members of the public are aware of assessment reliability issues, whether they perceive a need to be informed of reliability and, if so, the means by which such information should be delivered. The study takes a qualitative approach to the research questions and focuses on gauging the diversity of public opinion rather than quantifying responses among a representative sample. It forms the first part of Strand 3 of Ofqual's Reliability of Results research programme; the other Strand 3 project being a questionnaire survey which may build on the findings of this study.

3 RESEARCH AIMS

In order to explore the diversity of public opinion regarding the reliability of assessment results, this study was designed around five main aims:

1) explore the degree to which members of the public trust that grades or other assessment outcomes are accurate measures of performance or ability (i.e. their perceptions of assessment reliability),
2) gauge the degree to which members of the public are familiar with and accepting of human error,
3) gauge the degree to which members of the public are familiar with and accepting of measurement inaccuracy,
4) determine whether members of the public want to, or believe that they should, be (better) informed of reliability issues and, if so,
5) explore their perceptions of the best means of reporting reliability issues.

A focus group study was designed to address all five aims. The findings of the study are presented below, using each of the above aims to cluster particular themes in the participants' narratives. The main findings are also presented as hypotheses in order to generate ideas for possible items to be included in Ofqual's other Strand 3 work, the questionnaire survey.

4 METHOD

The chief methodological consideration for this study was how best to draw out opinions, beliefs and experiences on a topic that may not be familiar, understood, or seen as

particularly relevant to the participants' lives. A focus group methodology was used for the following reasons:

Reliability is a difficult concept to grasp (Ipsos MORI, 2009; Newton, 2005) and it was unlikely that members of the public would have a sufficient understanding of the issues to sustain a one-to-one interview. In contrast, focus groups lead to diverse and content-rich discussions as the interaction between participants can trigger the sharing of opinions, beliefs, memories, doubts and ideas (Kitzinger, 1994).

Focus groups are participant-led rather than researcher-led, which creates a collaborative rather than a researcher-dominant dynamic (Kitzinger, 1994). It was envisaged that this would encourage participants to define for themselves a possible way forward in terms of publishing reliability information.

Focus groups are an efficient means of collecting a large corpus of qualitative data in order to assess the variance or homogeneity of public opinion.

During focus groups it is possible to observe and explore what might influence opinion change. This was useful for hypothesising about whether the publishing of reliability information may increase or undermine public confidence in assessment outcomes.

Focus groups are often used as a precursor to quantitative work and the data are particularly suitable for developing hypotheses on topics about which little is known (e.g. Flores and Alonso, 1995; Morgan, 1997), as is required for the Ofqual Strand 3 questionnaire survey.

A focus group approach can also, to some degree, mediate the influence of researcher bias as the focus is on interactions between participants rather than the participants and the researcher (Krueger and Casey, 2000). In qualitative research bias can be interpreted as a tendency to gather, interpret and disseminate findings that are compatible with the researcher's world view, political attitudes, organisational affiliations, beliefs, priorities, or motivations (Hammersley and Gomm, 1997). As awarding body employees, the researchers were conscious that their insight into reliability issues could shape the questions asked and the conclusions drawn. The researchers examined and requested peer reviews of the research materials (the focus group schedule, prompts and vignettes) to uncover any potential sources of bias, and reflected on their interactions with participants. Qualitative research – like quantitative and experimental research - is filtered through the researcher and cannot be entirely free of bias (Hammersley and Gomm, 1997). However, the researchers were aware of their unique perspectives and were committed to ensuring that the processes of data collection and analysis were as free as possible from any unintentional sources of bias.

It should also be noted that qualitative research tends to draw upon smaller samples than quantitative research, and that the participants may have had their own reasons for wishing to take part. The views expressed in this report may not be representative of those of other direct and indirect qualification users. Some caution should be applied in generalising the findings of this study to similar groups or, indeed, the general public.

Ten focus groups were conducted, each comprising between five and ten participants and lasting approximately 90 minutes.

4.1 Operationalising reliability

Reliability refers to the dependability, trustworthiness and reproducibility of assessment outcomes; the extent to which a score or grade is free from random and systematic error and thus can be considered an accurate measure of an individual's knowledge, skill or ability (e.g. Nunnally and Bernstein, 1994). There are two main domains of reliability, both of which have several facets that tend to be measured statistically:

Test reliability – the degree to which the items in a test measure a coherent construct (internal consistency); the consistency of results when compared against those of a test designed to measure the same construct (parallel or equivalent forms), or the same test taken at a different time (test-retest reliability). Test reliability can also be influenced by test length, test environment and occasion, question type and quality, and the extent to which a test comprises a representative sample of the total universe of all possible test items or skill constructs. Further, these facets of test reliability interact with the ability profile of the examinees, with lower reliability coefficients typically observed among groups of similar ability, and higher coefficients among groups with greater variance in ability (e.g. Linn, 1993; Nunnally and Bernstein, 1994).

Examiner reliability – the degree to which an examiner has awarded a candidate their 'true' score compared with another candidate that they have assessed (consistent application of the mark scheme; intra-rater reliability), or compared against the judgement made by a senior examiner (inter-rater reliability based on a hierarchical approach to marking), or fellow examiner (inter-rater reliability based on a consensus approach to marking) (e.g. Baird, Greatorex and Bell, 2004; Meadows and Baird 2006; Linn, 1993; Nunnally and Bernstein, 1994).

Underpinning the various forms of reliability is the notion of assessment error. Newton (2005) proposes that assessment error consists of human error and measurement inaccuracy; terms that are somewhat conceptually indistinct and may be perceived as being overlapping. Using Newton's (2005) typology, human errors are associated with poor design or execution of assessment processes. These are usually categorised as occasion-, test-, marking- or grading-related and might include miscalculating a candidate's total score, failing to mark the whole script or misinterpretation of the mark scheme (Nunnally and Bernstein, 1994). Typically, these are avoidable errors, and of which the public can be relatively forgiving (dependent upon the scale and impact of the error) (Ipsos MORI, 2009).

In contrast, measurement inaccuracy is argued to be inherent to, and unavoidable in the educational assessment process (Newton, 2005). It is a more opaque and difficult to grasp concept and is therefore less familiar and understood (Newton, 2005). The term measurement inaccuracy is used to refer to the imprecision of assessment outcomes (Newton, 2005). Imprecision occurs as a result of the systematic and random occurrences that impact on candidates' assessment experiences and performances – these may include the candidate feeling ill, not being able to interpret an exam question, or a too hot or too cold examination room (Nunnally and Bernstein, 1994). Random errors and measurement inaccuracy are randomly distributed across candidates and cannot be predicted or controlled (Nunnally and Bernstein, 1994). Together with other occasion- test- marking- or grading-related errors, random errors and measurement inaccuracy can explain why a candidate might gain different outcomes on a number of different tests measuring the same construct or knowledge domain.

It was necessary for the focus group participants to engage with the concepts of human error and measurement inaccuracy in order to discuss their perceptions of the reliability of results and the issue of publishing reliability information. However, it was considered unlikely that participants would be familiar and able to engage with formal conceptions of reliability and measurement inaccuracy and relate them to their own experiences. Instead, it was envisaged that active discussion would depend on using non-statistical, non-technical terms and a more everyday definition of reliability.

Nunnally and Bernstein (1994) offer several definitions of reliability, with each definition emphasising a different aspect of reliability: usually the reproducibility, accuracy, consistency, or stability of measurements over time. Rather than use the term 'reliability', this study focused on Nunnally and Bernstein's (1994) notions of accuracy and reproducibility. It was considered that they would trigger a more concrete understanding of the issues to be discussed. 'Reliability' was therefore conceptualised as:

the extent to which we trust that our assessment outcomes are accurate and reproducible in different circumstances.

Although this definition was not shared explicitly with participants, it was useful for structuring the focus groups. Reliability (the 'accuracy' and 'reproducibility' of outcomes) was introduced as the product of the combination of candidate ability, human error and measurement inaccuracy.

Previous research suggests that public understanding of awarding body processes and technical assessment concepts (such as examination error) is understandably hazy (Ipsos MORI, 2009; Newton, 2005; Taylor, 2007). The language of assessment and the processes of awarding bodies are not familiar or widely understood. To overcome this, to clearly operationalise the term 'reliability' and to focus the minds of the participants on the pertinent issues, a set of vignettes were developed (see Appendix 1). Vignettes are often used to contextualise a discussion, introduce topics that are difficult to understand or sensitive, and to allow participants to talk about the circumstances, actions or predicaments of a fictional person rather than themselves. The vignettes used in the focus groups were very short stories or scenarios involving fictional characters in specific dilemmas which were related to the research context and relevant to the lives and educational experiences of the participants (Barter and Renold, 1999). The participants were asked to read a vignette at a specific point during the focus group and discuss their responses to it. Each vignette targeted a specific reliability-related scenario in the context of teacher assessment, public examinations or vocational qualifications, and was used with the focus group participants likely to be most familiar with the context of the vignette. The vignettes were used only as a trigger to facilitate discussion about participants' beliefs and attitudes towards reliability issues – their responses to each vignette are not therefore discussed explicitly, but were subsumed within the thematic analysis below.

4.2 Sampling and the participants

Krueger and Casey (2000) state that 12 focus groups is usually a sufficient number to reach saturation point; the point at which it is less likely that new ideas and opinions will emerge. Given the limited time frame for this study, ten focus groups was considered achievable. The target number of participants per group was six, although eight participants were typically recruited to allow for some attrition. Where there was more doubt about the participants'

commitment to attend the session, as many as ten participants were recruited. In the event, the groups experienced very low attrition rates, with only two participants choosing not to attend. The lack of attrition was attributed to the incentive[1] and the fact that the focus groups were held in locations and at times that were convenient for the participants. For example, the NHS and teacher groups were conducted at the participants' place of work at the end of the working day, while the PGCE groups were conducted during the students' extended lunch break in a PGCE training classroom.

Qualitative research tends not to be concerned with issues of representativeness; instead the focus is on exploring the diversity of opinions, attitudes and experiences. The sample was selected to represent the views of the five direct and indirect educational-qualification user groups listed below. In order to simplify the data collection process, all participants were located in the Greater Manchester area.

1) **Students:** two groups of PGCE students from the University of Manchester.
2) **Employers:** two groups of employees responsible for recruitment within the Central Manchester University Hospitals NHS Foundation Trust.
3) **Employees:** two groups of employees of the Central Manchester University Hospitals NHS Foundation Trust.
4) **Job-seekers:** two groups of individuals currently unemployed and seeking work.
5) **Teachers:** one group of secondary school teachers (from a grammar school in Greater Manchester) and one group of primary school teachers (from a primary school in Oldham).

PGCE students were included in the study as their grade profiles (including GCSE Maths and English) will have been central to gaining access to teacher training courses. As trainee teachers it was also considered that they may be more likely to engage with issues of reliability, teacher assessment and national assessment than students of other sectors and disciplines. It is worth noting however, that assessment is rarely a dominant theme on PGCE course outlines - a willingness to engage with the issues may not necessarily go hand-in-hand with a sound understanding of the issues.

Similarly, job-seekers were included as it was considered that they would have a unique perspective on issues of assessment reliability given the centrality of their qualifications profile to their job search. The job-seeker groups proved the most difficult to recruit. Following several failed attempts to recruit via job centres[2], these participants were recruited using the snowball sampling method. This entails making one or more initial contacts and following up any leads suggested by the contacts. The method is commonly used for qualitative research as it is informal and useful for accessing hard-to-reach participants (Krueger and Casey, 2000). In this case, announcements were made at the end of one NHS employee and one NHS employer focus group that job-seekers were required for future focus groups. The NHS participants were given fliers to give to anyone they considered to be a job-seeker. As a consequence of using this recruitment method, the job-seeker participants tended to be

[1] Participants in the teacher, employee and employer groups were given a £50 Marks & Spencer voucher. The PGCE students and job-seekers were given a £50 Tesco's voucher.
[2] Job centres were unable to grant the researchers direct access to their clients, or allow the researchers to approach clients on job centre premises. Posters and fliers were given to two job centres with an agreement that these would be distributed to clients attending 'Back to Work' courses. It is unknown whether this occurred as none of the participants were clients of the job centres. The 'Back to Work' course instructors suggested that the participation rate was likely to be low as their long-term unemployed clients tend to lack the skills required to take part in a group discussion.

recent graduates (Further Education, under-graduate or post-graduate) who were friends or family of the NHS participants.

The four groups of NHS employees/employers were selected on the basis that comparisons are often made between medicine and education in terms of the need for accountability and transparency. It was envisaged that medical practitioners would be able to comment upon their educational experiences as well as compare measurement error and the potential for grade misclassification with misdiagnoses and other medical errors. For the purposes of this study 'employers' were defined as individuals with a responsibility for recruitment of behalf of their institution (the NHS) and their department.

Parents were not included as a distinct group as it was anticipated that the views of parents would be represented across the focus groups. This proved to be the case, and many participants were able to recount experiences of their own and those of their children, and discuss their attitudes towards reliability issues from their perspective as a parent and more generally.

Krueger and Casey (2000) note that it is typically very difficult to recruit participants for focus groups, and particularly difficult to recruit male participants for any kind of social research. Potential participants may be uncertain about the format, reluctant to interact with strangers, or anxious about speaking in a group context. This can be exacerbated if the topic is complex, sensitive, or unfamiliar – which clearly may be the case with assessment reliability. The following table shows that a sufficient number of participants were recruited to each group, and that with the exception of the primary school teachers, each group had adequate male representation.

Table 1. Focus group composition and method of recruitment.

Cohort	Group 1	Group 2	Total (n=74)	Recruitment method
Employees	2 male, 8 female	3 male, 5 female	5 male, 13 female	NHS intranet announcement
Employers	5 male, 4 female	1 male, 5 female	6 male, 9 female	NHS intranet announcement
Job-seekers	3 male, 4 female	3 male, 2 female	6 male, 6 female	Snowball
PGCE Students	3 male, 5 female	5 male, 4 female	8 male, 9 female	Request via PGCE Course Leader
Teachers	Secondary 3 male, 3 female	Primary 6 female	3 male, 9 female	Direct contact with schools

4.3 Sampling limitations

As the aim was to recruit sufficient participants within a restricted amount of time, very few selection criteria were applied. It was important that potential participants could demonstrate their affiliation to each particular social grouping, but this was largely inherent to the process as participants were often contacted at their place of work or study. In addition, when groups were over-subscribed (particularly the case with the NHS groups) participants were selected with a view to achieving diversity in terms of age and gender. Missing from this, however, was any attempt to ensure representation from ethnic minority groups. Just five of all the participants represented groups other than White-British. MORI (2003) purports that a lack, or loss, of trust in public institutions and agencies is more pronounced among black and minority ethnic communities. This report is unable to comment upon the reliability of this claim in relation to assessment agencies, but it may be prudent to ensure there is sufficient ethnic minority representation in the subsequent questionnaire survey component of Strand 3 of the reliability project.

Additionally, all participants were well educated and thus did not represent a diverse range of educational attainment. The majority of participants had completed under-graduate degrees; some had, or were undertaking, post-graduate qualifications such as PGCE, a Masters degree or PhD, and many of the NHS participants held advanced professional qualifications (required for practice as a dietician or radiographer, for example). Although it was not the intention to recruit such well educated participants, in retrospect there were several advantages to this. In particular, the participants had:

- extensive experience of preparing for and taking examinations,
- an awareness of the existence of awarding bodies and their responsibilities (although there was still considerable doubt about this issue),
- experiences of human error in the examination process, and requests for re-marks and appeals,
- an understanding of the examiner's role, and
- experience of using their qualifications to compete for educational and job opportunities (and hence the importance of reliability).

To varying degrees, the participants were also able to grasp the concept of measurement error and discuss their responses to it. The focus group discussions may not have been as fruitful with individuals who were less familiar with the process of educational assessment. As this study serves as the groundwork for the Strand 3 questionnaire survey, it therefore appears to have been a benefit to have canvassed the opinions of well educated and informed individuals.

4.4 Focus group format

A focus group schedule was designed to cover the following objectives:

- introduce the participants to each other (not required with the teacher groups) and the facilitator,
- focus the participants' minds on educational assessment,
- explore their good and bad experiences of assessment,
- gauge their level of trust in assessment results, awarding bodies and examiners,
- explore their experiences of, or introduce them to the issue of human error in the assessment process,

- introduce them to the concept of measurement inaccuracy,
- assess whether introducing them to issues of human error and measurement inaccuracy influenced their level of trust, and
- explore if and how issues of reliability should be best communicated to the public.

Given the complexity of the issues, it was essential that these issues were introduced in a structured and accessible way. Table 2 shows a rough timeline of each focus group and the aims of each section. All groups followed the same schedule, with the exception of the secondary school teachers. This group was particularly knowledgeable about human error and measurement inaccuracy and it was decided during the course of the discussion that the vignettes were not required.

To maintain consistency, each focus group was conducted by the same facilitator and, with two exceptions, the same observer. The facilitator has a critical role and, depending on the level of trust they foster among the participants, can encourage or inhibit open discussion (Gibbs, 1999). Consequently, an experienced facilitator-observer team and consistent approach to conducting each focus group was essential. The observer took notes to identify each of the participants, note emerging themes, note the time of any particularly interesting comments or discussions, and monitor the audio equipment. The facilitator steered the discussion by providing open-ended questions and verbal prompts, ensured the active participation of all group members, and moderated the influence of any dominant personalities.

All participants signed a consent form which informed them of their right to withdraw from the research, the fact that all data would be treated as confidential, and the intended uses of the data. With the participants' permission, all discussions were audio recorded.

Table 2. Timeline and aims of each focus group discussion.

Time (minutes)	Question/activity	Aim
0-5	Facilitator's introduction to study	Explain aims of study; rules for conduct of focus group; confidentiality
5-15	Ice-breaker question: 'Tell us about your most memorable examination experience, good or bad'	'Warm-up' for participants; focus participants' minds on assessment; gather assessment experiences that could be referred to later in the discussion as examples of human and random error
15-25	Assessment error: 'Have you ever been shocked or pleasantly surprised by an assessment result; how do you think that happened?'	Begin to introduce concept of error; how do we explain discrepancies between predicted and achieved grades, or teacher assessment and examination results – what could have happened to explain discrepancy (e.g. marking error, poor performance on the day, 'got lucky' with exam questions etc.)?
25-35	Trust: 'Thinking about the exams you took at school, to what degree did you get the results you deserved?'	Establish level of trust in accuracy of assessment results
35-50	Error: Vignette designed to encourage discussion of error in either the context of grades, vocational qualifications or teacher assessment	Discuss potential causes of error experienced by the fictional vignette characters; who or what explains 'undeserved' results?
50-65	Comparing reliability across types of assessment: Vignette designed to compare trust in academic and vocational, or teacher assessment and national qualification	Establish whether participants trust one mode of assessment over another (different vignettes used for different groups e.g. NHS groups asked to compare vocational and academic; teachers asked to compare teacher assessment and public examinations)
65-75	Introduce concept of measurement inaccuracy	Discuss the acceptability of some degree of unreliability; gauge whether this impacts on participants' trust in the accuracy and reproducibility of assessment results
75-90	Communicating reliability to the public: how and what information	Assess whether the public want to, or feel they should, be informed of reliability issues and if so, discuss the best means of doing so

4.5 Data analysis

The data were analysed using framework analysis (Ritchie and Spencer, 1994) which encourages data collection and analysis to be conducted concurrently. The aim was to organise the data into discrete themes that cut across the different focus groups, and identify quotes to exemplify each theme. The data analysis process consisted of the following overlapping stages.

Stage 1: Data collection and familiarisation
The observer and facilitator took field notes during each discussion and spent approximately 10 minutes writing a summary of the discussion at the close of the focus group. They then discussed their notes and their perceptions of the focus group more generally, noting emerging themes, different conceptions of reliability, and on which themes the participants were particularly strong or weak. These comments were referred to repeatedly during the latter stages as pointers as to which groups and participants commented on a particular topic.

Stage 2: Familiarisation and searching
The recordings were transcribed as soon as possible after each focus group by a secretarial agency. The first reading was conducted within three days of the focus group while the content was still fresh in the minds of the researchers.

Stage 3: Searching, identifying thematic framework and mapping
The transcripts were imported to NVivo (QSR International, 2008), a qualitative data analysis software package. With each subsequent reading irrelevant data were deleted and emerging themes built upon, combined or eliminated depending on the weight of evidence to support each theme across the groups. The themes were not pre-determined, but emerged from the participants' narratives. The final themes were defined by their extensiveness – the degree to which similar comments were made by several people – or by the participants' strength of feeling for a particular topic.

Stage 4: Mapping and interpretation
The facilitator conducted the analysis and met with the observer several times to discuss the emerging thematic framework. Themes were negotiated to ensure their validity. At the final analysis stage the researchers selected the quotes that best illustrated each theme for inclusion in the report.

As each of the five qualification-user sectors was represented by only two groups of participants, it could be misleading to discuss findings by group, so findings are presented below by theme rather than by qualification-user group. However, there are some exceptions to this where the opinions of one group differed to others. Any significant differences are noted where appropriate.

5 DISCUSSION OF FINDINGS

The ten focus groups produced a large corpus of qualitative data containing a number of complex and overlapping threads of discussion. In order to simplify the discussion of findings, the five aims of the study are used as umbrella terms to group the themes. Themes are used to explain or contextualise the summary of the findings relating to each research aim.

5.1 AIM 1: PUBLIC PERCEPTIONS OF RELIABILITY

With the exception of the secondary school teachers, participants' awareness of assessment reliability was limited. Most participants had not previously thought about the reliability of their assessment outcomes and instead had an implicit trust that the examination process awarded them the grades that they deserved. On the whole, the participants believed that examination personnel were educational experts not prone to making errors (with the exception of the occasional human error). In cases where participants felt there was a discrepancy between their predicted and achieved grade, or their perception of their performance and achieved

grade, the participants tended to blame themselves rather than attribute it to any unreliability or error in the process (except in the case of a better than expected result which was more likely to be attributed to 'error'). Some participants had experience of re-marks and appeals. This appeared to make them more likely to question potentially suspect outcomes, but on the whole did not appear to diminish their trust in assessment outcomes to any significant degree.

Five themes illustrate the findings relating to public perceptions of reliability: trust in the assessment process, trust in examiners, trust across different modes of assessment, when things go wrong: attributing 'blame', and 'you get what you deserve'.

5.1.1 Trust in the assessment process

The issue of trust in the assessment process produced the most significant difference in opinion between the teachers (secondary school only) and all other groups[3]. As the following quotes suggest, the non-teaching participants and, to a lesser extent, the primary school teachers, tended to have an almost blind faith in the ability of the assessment process to award accurate grades.

> *I* [did my exams with] *the Joint Matriculation Board and they're this big power and you just respect the authority; I didn't doubt my results at all.* (Primary school teacher, female).

> *It's that kind of underlying trust that you have in the examination board that they're going to get it right and that all the systems will be put in place to make sure you are getting the grade that you deserve.* (NHS employee, female).

> *It comes down to the fact that you've got to hope that it's fairly evaluated, assessed and graded. It comes back to having trust in the examination system isn't it? We trust the examination boards without question - foolishly I expect.* (NHS employee, female).

> *It's what you're ingrained to trust. You're ingrained to trust because that's the system that's always been established, that a written exam is the best way.* (NHS employee, male).

> *There's a significant amount of trust in the examination boards to get our results right.* (PGCE student, male).

A small number of participants commented, however, that the era of blind faith in large institutions may be coming to an end, and that, as a society, people are more likely to question and challenge. This view mirrors that of O'Neill (2002) who suggests that a 'loss of trust' characterises the contemporary social world, and that people are increasingly mistrustful of a range of private and public agencies. However, this was an unusual view among the participants; the majority suggested that they had had no reason to question the reliability of their outcomes, or the assessment process as a whole. It may be suggested that, thus far, assessment has escaped the kind of sustained public scrutiny to which other social systems – the medical, judicial, and social work systems for example – have been subjected. As noted earlier, the spotlight appears to fall on assessment only sporadically and to some degree,

[3] On this issue, the opinions of the primary school teachers were more in line with the non-teaching groups.

superficially. The sentiments expressed by the following participants were therefore somewhat unusual:

> *I think people have less faith in institutions generally now than they did. I don't think it's specifically exams. It's just a general thing. People are more inclined to doubt an actual institution, be it an examiner or what have you, than they would have done, say, ten, twenty years ago.* (Job-seeker, female).

> *There's more of a culture of that, now, though, isn't there, in questioning these sort of things? I think that sort of culture didn't really exist so much* [when I took exams]. *The people who raised A-level queries when I was at college had been predicted a higher grade. But it was unusual, they had to go through like a real rigmarole. It's definitely more a culture of questioning how rigorously these things are being done now.* (PGCE student, female).

> *I think the culture may be changing. I think we're more of a challenging culture than we were. When I was younger you did what you were told or you believed what you were told and that was it. You never questioned and you never challenged. But it's different now. You tend to challenge more.* (NHS employer, female).

Perhaps inevitably the secondary school teachers had a different perspective of trust in the reliability of assessment outcomes and of the reliability of examiners' marking. The teachers were generally satisfied with the reliability of A-level outcomes, believing that they often met their expectations. As two teachers suggested:

> *Our accuracy at predicting A-levels is stunning. It's to within a per cent.* (Secondary school teacher, male).

> *I think at A level there are certainly anomalies, but by and large at A-level* [the outcomes] *aren't so much of a lottery.* (Secondary school teacher, female)

They suggested, however, that GCSE outcomes could be more variable, unpredictable and, to an extent, unreliable, as outlined by the following participant:

> *We've got all those wonderful AS scores which are already pointing very strongly in a particular direction. GCSE is 'take your pick', almost, and 'pot luck' and various other expressions of randomness.* (Secondary school teacher, male).

The secondary school teachers had experience of challenging candidates' results, and these experiences appeared to undermine their trust in the reliability of outcomes. There are several reasons why teachers might have more faith in the accuracy of A-level than GCSE grades, including that the basis for statistical predictions (used to support awarders in selecting boundaries) at A-level is arguably more robust than that used for GCSE. The perceptions of the secondary school teachers are also supported by evidence relating to the reliability of GCSE and A-level outcomes. Dhillon (2003) reported that there was an unprecedented degree of accuracy between teachers' estimates and the first year outcomes of the Curriculum 2000 A-levels[4], but the same has not always been found for teacher estimates for GCSE (e.g. Delap 1994). Delap (1995) and Murphy (1979) also report that teachers tend to

[4] The same level of accuracy was not found for AS level – only A2 – and it is unknown whether this relationship has changed over the history of the Curriculum 2000 specifications.

over-estimate the performances of their candidates, meaning that they will most likely be dissatisfied when achieved grades fail to meet their estimates.

The following participant offers another explanation for unanticipated results; that candidates may provide responses that challenge the examiner. Research has found that candidates from independent and selective centres (like this teacher's students) are more likely to provide unusual responses that push the limits of the mark schemes (Pinot de Moira, 2005). Awarding bodies strive to mark all responses as reliably as possible by producing full and detailed mark schemes, but unusual and unanticipated responses represent a challenge for examiners. Ahmed and Pollitt (2008) identify such responses as a threat to validity and as evidence that the writer of the question and mark scheme has not anticipated all possible means of interpreting and answering the question.

> *We've had some run-ins with exam boards in the past, haven't we? We've gone in with the jackboots on, really, because we've felt the kids have been unfairly done to. We've had results go both ways, but we've had a couple of major ones go in our favour where the work just wasn't examined effectively. I'd got this funny feeling that the kids actually knew more about what they were doing than the person marking it.* (Secondary school teacher, male).

At times, the language used by the secondary school teachers suggested that the relationship between teachers, awarding bodies and examiners was fraught, and underpinned by a level of distrust. The teachers' level of frustration is evident in the following two-person exchange and quote:

Male: "I think we won a couple of battles..."
Female: "We have. We have."
Male: "We have. We won some quite major battles with GCSE."
Female: "Oh yes, huge."
Male: "They were putting kids' marks up by 20 percent, and that's horrific."

> *I have a teacher pulling her hair out at the moment because of a paper that she was convinced was far too hard. She contacted six other schools like ours in the area. They all agreed* [the paper was too hard]. *Then they worked out, 'Oh, it's been set by him. It's that mad examiner who's been around for years and he's just gone off on one this year'. So six schools of this type in this area are doing that syllabus and the kids are penalised by that mad examiner. They can do nothing. That mad examiner has an effect on every child in the country doing it.* (Secondary school teacher, male).

This participant felt strongly about the existence and influence of the 'mad examiner', but essentially they are referring to the difficulty faced by senior examiners in producing papers of consistent demand year on year. The work of Pollitt in combination with Ahmed and others (e.g. Pollitt, Ahmed, Baird, Tognolini and Davidson, 2007) to develop a systematic approach to writing questions of known demand demonstrates how difficult this is to achieve in practice. Producing questions of known and varying demand depends on a thorough understanding of question construction and how the parts of a question can be manipulated to vary question demand (e.g. Crisp, Sweiry, Ahmed and Pollitt, 2008).

The participant's comment also suggests that teachers may be unaware of how the awarding process can mitigate the influence of fluctuating demands over successive years. The secondary school teachers suggest below that they believe any initiative to improve public understanding of assessment reliability should start with students and teachers. It may be the case that an enhanced understanding of awarding processes may give teachers a greater insight into the relationships between question paper difficulty, demand and standards.

The difference in attitude between the secondary school teachers and all other groups was striking. The insight, experience and knowledge of the secondary school teachers made them ready to question and challenge assessment outcomes (more so at GCSE than A-level). In contrast, the participants of the other groups had rarely thought about or had reason to question their assessment outcomes, and had an implicit trust in assessment systems to deliver the 'true' results. Crucially, these views were expressed prior to discussing the existence of measurement inaccuracy. Once participants had been encouraged to identify and discuss all potential sources of measurement error, they tended to suggest that they would be more cautious in trusting their outcomes if they failed to meet their expectations (or those of their children).

5.1.2 Trust in examiners

For many participants their trust in the assessment system was underpinned by a belief in the professionalism and knowledge of the examiners. As the following quotes suggest, however, this belief is somewhat subconscious and rarely articulated or questioned.

> *I think I've always tended to trust the examiner as to what he or she thinks of my knowledge. I've never really questioned it. They're the subject experts. At O-level I considered myself as a kid who was dabbling in Chemistry, Biology and Physics, not an expert in any shape or form so I just trusted the examiner.* (PGCE student, male).

> *I think probably when I was a teenager I'd think that some really clever person was marking* [my script]*; somebody who was a real authority and knew what they were doing. I'm not quite as naive now, you know, but I would think then, that this is a person who really knows their stuff.* (Primary school teacher, female).

Many participants were aware of the pressures that examiners operate under, either because they had friends who had worked as examiners or simply that they could envisage the weight of an examiner's workload during the summer period. As the following quotes suggest, an awareness of workload pressure and the potential for error did not necessarily undermine their trust in the system.

> *You just kind of hope and trust that they are doing it properly… you know, they're not doing them all in one go and getting really tired and they're doing it properly sort of thing and using the criteria. You've just got to hope that they are really.* (Job-seeker, female).

> *I don't know but I'm guessing that they're marking these papers in the evening and they've probably already worked ten hours in the day and they're going to mark for four hours in the evening and they've probably got a glass of wine in front of them or something. Who knows, who knows?* (Job-seeker, female).

> *But isn't there a certain standard that they have to reach? I mean, surely it's not grabbing people off the streets and saying, 'Mark a paper'. They've got to be at a certain level and there are certain standards that they've got to meet in order to mark that paper – or you'd hope!* (NHS employer, female).

Similarly, the participants recognised that the marking of some subjects may be more subjective than others. This also did little to diminish their trust in examiners to deliver the 'true' marks. As the following quotes show, the participants tended to see subjectivity as inherent to the assessment process rather than as a source of 'error'. In this sense the participants appeared to prioritise test validity over reliability.

> *It depends what the paper is. Say it's maths, there is really only one answer isn't there? But if it's something like English and you have to give an opinion then maybe if the marker agrees with your opinion you might get a higher mark; if they disagree then you won't.* (Job-seeker, male).

> *If you've got to write about an opinion; it's going to be someone else's opinion if they agree with you or not. There's no right or wrong answer with an opinion is there?* (Job-seeker, female).

> *Maths is easy because it's one and one is two. A lot of other subjects are very difficult to interpret, like History; it's very difficult to interpret.* (NHS employee, male).

> *Science is easy* [to mark/interpret]*, but other subjects, essay subjects like English and History, are more open to interpretation.* (PGCE student, male).

Many of the participants showed a good awareness of the different demands that different examinations place on examiners. It can be inferred from their comments that they quite rightly identified that reliability is variable according to the subject matter content and, by extension, the question types used to assess candidates. The marking reliability literature provides numerous examples of studies showing that closed-response questions (such as multiple choice questions), which are usually accompanied by tightly defined mark schemes, produce consistently higher reliability coefficients than open-response questions such as essay questions (see Meadows and Billington (2005) for a review of the marking reliability literature). Without necessarily articulating the technicalities of how some subject areas may be more reliably marked than others, the participants appeared to accept that some subjectivity was an inevitable part of the assessment process, and did not see this as an unacceptable form of error.

5.1.3 Trust across different modes of assessment

The most fruitful discussions about trust in different modes of assessment took place with PGCE students and the primary and secondary school teachers. The PGCE students were split between favouring coursework and teacher assessment, and public examinations. Some participants suggested that examinations provide only a 'snap-shot' of performance, whereas coursework and teacher assessments are based on a more rounded and thus robust evidence base, as the following participants suggest:

> *I'd be more inclined to trust a teacher's estimate. If they have worked with these students for the two years up to GCSE on a day-to-day basis, they've got past tests and homework to work from and presumably end of year exams that the school would administer. I'd be surprised if their predicted grades were that much out.* (PGCE student, male).

> *I think nowadays we're moving over to more coursework. I did O-levels and it was all or nothing on the day. Two hours, it was a life-changing thing. Coursework is a much fairer way of doing it because you've got the best of both worlds. You can show what you can do and you've got your teacher who knows you and who knows your abilities.* (PGCE student, male).

Other participants felt more confident in the reliability of public examination outcomes, believing that the scale of the assessment system and professionalism of examination personnel created reliable indicators of performance. The participants quoted below framed their preference for exams in the context of the potential unreliability of teacher assessment; identifying for example that teachers might over-estimate the performances of their students, as indeed has been found to be the case (e.g. Delap, 1995; Murphy, 1979).

> *I don't know if it's just because it's been embedded in me or just because that's how I've gone through the education system. I would trust the exam result and think that maybe the teacher was favouring me a little bit.* (PGCE student, female).

> *I'd trust the exam board* [result] *simply because the teacher is just one person whereas the exam board is going to have systems in place and is going to check and double check. There will be an entire group, say a governing body, who decides the criteria. You have more faith because it's a larger body.* (PGCE student, male).

> *I trust the exam board more; they're seeing so many more than what the class teacher sees.* [A teacher] *might see twenty-eight pupils. The exam board sees thousands and thousands. I like to think that they know more what they are looking for than what the class teacher does. They're taking the bias out of the equation, saying "Yes, it's right" or "No, it's wrong". It's taking out the human element.* (PGCE student, male).

The primary school teachers had considerable experience of teacher assessment and had a strong belief in the reliability of their judgements. Again, the teacher participants appeared to articulate in layperson terms some of the issues that are of interest to assessment experts. Like the second participant below, Wiliam (2007) suggests that teacher assessment can enhance test validity as it offers the opportunity to sample more of the curriculum. It is also a means of enhancing reliability as it lengthens the test, which Wiliam (2007) argues is currently the only accepted – although not unproblematic - means of increasing reliability. The second participant appears to recognise that there are limitations to what can be inferred from test results.

> *Well,* [teacher assessment is] *not just accurate, but honest. We're quite confident in what we're doing* [when we assess pupils]. (Primary school teacher, female).

> *We have confidence that we can back up our teacher assessment with evidence.*

> *The test results that we also use to make up the picture becomes less and less important because it's just one snapshot, one simple day of that child's life.* (Primary school teacher, female).

The secondary school teachers were also confident in their ability to make reliable judgements about performance. They conveyed a sense of frustration, however, that their judgements of candidates' internally-assessed coursework could be undermined by moderators. As the following quotes suggest, they believed the quality of moderators and moderation was variable:

> *The moderators don't have experience in terms of applying the objectives.* (Secondary school teacher, female).

> *You can't guarantee – our experience is that you can't guarantee – quality of examiners year on year. Also, you can't guarantee consistency on marking moderation is the same year on year.* (Secondary school teacher, male).

In the light of some unsatisfactory experiences with the quality of marking and moderation, the secondary school teachers were more inclined to choose specifications with internally marked coursework components, and to become moderators themselves. They suggested that this served as 'insurance' that their candidates would get the 'right' outcomes:

> *You look for the* [specification] *that's got coursework which we can mark ourselves and apply criteria ourselves in a consistent way.* (Secondary school teacher, male).

> *Two of us in the department are head of moderators and the reasons for that are, one, to gain understanding and the other is that it makes it harder for another moderator to come in here and say, 'You are wrong', when we've got two moderators that have already marked it. It's insurance, isn't it? The number of moderators that moderate because they feel that at some stage in their career they've been done over is quite high.* (Secondary school teacher, male).

> *With the new AS... we'd gone through standardisation and sent off the marks, happy that we'd got it right. Then we went out to other centres and looked at how they moderated there and thought, 'Christ, we've under-marked ours'. It's all very well if we're accurate, but if everybody else is marking higher than us and moderators go in and don't re-standardise it, then we're going to do our pupils a disservice.* (Secondary school teacher, male).

5.1.4 When things go wrong: attributing 'blame'

Perhaps not surprisingly, participants were mostly concerned about outcomes that were poorer than expected; participants were not concerned about or likely to question results that were better than expected. Interestingly, while getting a better grade than expected caused uncertainty and confusion, participants seemed more likely to attribute this to error, rather than their performance, as the following participants suggest:

> *I've actually thought to myself, hold on, I'm sure I've been over marked in some examinations. I specifically remember in GCSE French I got an A and I thought that was completely a fluke because I barely spoke any French and I certainly didn't*

> *think that I did well in any of the examinations. And I thought, well, if it had been the other way then yes I would definitely query it because I got a mark below what I thought I would get. But I didn't question it at all even though I thought to myself I should be questioning this.* (PGCE student, male).

> *I got a B in Maths O-level and I was absolutely rubbish at maths. I got a B and I can't imagine how I got that. I don't know how I got that.* (NHS employer, female).

> *If* [the result] *is better than you expect, do you really mind if you've got a false grade? I don't think most people would. You'd take the money and run, wouldn't you?* (NHS employer, male).

The implicit trust that many participants had in the assessment process meant that they tended to blame themselves in the case of a poor result. Rarely did the participants question whether an error could have occurred, or whether they would have achieved a different outcome on a different day or with a different test, as the following participants suggest:

> *I had nobody to blame but myself* [for my poor result]. *I really, really didn't.* (Primary school teacher, female).

> *Teachers, especially school age, you're told it's you, that if you don't do the work you won't get the grade. You're not encouraged to question whether that was the right grade or whether you should have got higher, or whether you should appeal.* (NHS employee, female).

> *You'd just think well it's down to me, that grade is because I didn't perform well on the day or I didn't do enough study or the coursework was below standard. You wouldn't necessarily think well the examiner is at fault here. You'd look at yourself first.* (NHS employee, female)

> *When you sit an official exam you don't query it. I always felt responsible. If I'd done poorly on an exam, it would always be my fault for not revising hard enough.* (Primary school teacher, female).

The participants' perspectives on attributing blame were interesting and ran counter to what we might expect. Attribution theory is concerned with the methods by which individuals maintain a positive self-image and make sense of their successes and failures (Weiner, 1992). One of the key assumptions is that, to maintain a positive self-image, individuals explain failures with reference to external, environmental factors, and attribute their successes to their internal qualities. In applying attribution theory in an educational context, we would expect an individual to 'blame' a poor test result on poor teaching or bad luck, and attribute a good test result to their intelligence, skill or effort. The participants' comments mostly suggested the opposite; that an unexpected poor result was 'blamed' on internal factors such as a lack of effort, and unexpected good results were attributed to external factors such as error or luck. Unexpected results appeared to cause uncertainty and it may be the case that the participants lacked a sufficient understanding of the assessment process to explain their result to themselves.

5.1.5 You get what you deserve

The participants' trust in examiners and the assessment process, and the fact that participants mainly blamed themselves for poor outcomes, seemed to go hand-in-hand with the belief that, when it comes to examination results 'you get what you deserve'. When talking about their examination results generally (and not 'unexpected' results) the participants' views were consistent with the tenets of attribution theory (Weiner, 1992). As the following quotes suggest, their examples of examination success were attributed to internal factors including effort and intelligence:

> *My daughter was absolutely delighted and she got what she deserved because she worked hard. And my other daughter who's the very smart one; she didn't do nearly as much work as the second one but she got what she deserved because she's bright. So it does work out that way, you know, if you put the effort in you're going to get the marks that you deserve.* (Job-seeker, male).

> *Most people think that they do get what they deserve. If you put a lot of effort in you're probably going to get good marks; if you don't you probably won't. And that's the general rule of thumb; it seems to work quite well.* (Job-seeker, female).

The notion that 'you get what you deserve' illustrates the participants' belief that ability is the greatest, perhaps only, predictor of attainment, and their lack of awareness of measurement and reliability issues. For them, a grade is a direct correlate of ability and the amount of effort made in class and in preparation for an exam, and very little else. Interestingly, the secondary school teachers did not express anything approaching a similar view, suggesting that they appreciate that there is more to the assessment process than simply getting 'what you deserve'.

5.2 AIM 2: ACCEPTABILITY OF HUMAN ERROR

The participants were able to list a number of potential types of human error, either through direct experience or being able to envisage how errors may occur during the assessment process. The participants appeared relatively tolerant of human error, both globally and in relation to their own examination experiences. Given the scale of the assessment system they suggested that human errors were, to some degree, inevitable. The exception to this was typographical errors or errors in the distribution of examination materials, of which the participants were far less tolerant. The participants with children suggested that, in the case of an unexpectedly poor result, they would be more inclined to question whether an error had occurred, than they would have been for themselves – which may be a product of a more challenging society or of the parental role.

5.2.1 Understanding human error

A number of participants had either direct experience of human error or could appreciate how human error could occur. In general, even those who had experienced human error were relatively forgiving; those who had challenged their results tended to be satisfied that the re-mark or appeal process had resolved the error, and those who had not challenged their results had accepted their grades, flawed or otherwise, and perhaps with the feeling that they had not quite got what they had deserved.

The following quotes show the participants' awareness of some forms of human error:

My partner did some temp examiner work a few years ago for EdExcel. They were given these incentives to encourage them to mark quicker, and people who marked the most papers in a given amount of time were given the incentives. And he said how, you know, some people at the back of the room were falling asleep and stuff like that, at their computer as they were doing the marking. And it made me think how easily a clerical error could arise if people are either not paying attention or if they're trying to do it really quickly to try and get these incentives. (Job-seeker, female).

I suppose errors can happen if the marks aren't added up correctly. And sometimes, the exams get sent off, you know, and you could have a missing piece and lose components of the exam paper, you know, if you're using extra bits of paper. (NHS employee, female).

Presumably, essays are harder to assess. Multiple choice questions are either right or wrong. In a way, they can be read by a barcode reader and you just get the marks out. An hour-long essay can be perceived in many different ways. Somebody may say, 'Yes. This is spot-on'. Somebody else may say, 'This is a load of rubbish'. (NHS employer, male).

5.2.2 Tolerance towards human error

The following quotes show the general level of tolerance towards or acceptance of human error:

I've got a friend who does exam marking, and like, she's just a normal secondary teacher but she does some over the summer and she said she just had like piles of papers and she was just ploughing through them and I just think, if you're sat there, ploughing through loads of marking, you must make the odd little mistake. (Primary school teacher, female).

You get the odd error coming up in the system. I had a Spanish A-level exam that was marked and I think I got something like 45 UMS out of about 105. When we sent it back for re-marking, it came back as 86. So sometimes you do get some strange things. (PGCE student, female).

It would be unrealistic to think that so many exam papers can be marked each year without some margin of error. All of us can see how those things could happen and you take that into account when you get your grades or when you sit an exam, that that's what you're sort of buying into. It would be slightly unrealistic to expect everything to be 100 percent all the time. (Job-seeker, female).

I must have done 50, 60, 70 exams during the course of my career so far and the chances are that on one of them I will have got the wrong grade. Given the number of people and the amount of exams they do, chances are most people have probably had the wrong grade at some point I'd imagine. (PGCE student, male).

Interestingly, the participants were less tolerant of typographic errors or other errors associated with examination materials, which arguably have less of a direct impact on their assessment outcomes than other forms of error. This perhaps reflects the participants'

perceptions about what types of error are avoidable and which types are inherent to the process and more difficult to eliminate.

> *I don't think there's much excuse for it, though, if you think of the amount of text on most exam papers, if a few people check it through, surely you know some people would be able to pick up on it. It's not like if you've done a typo in a huge thesis or something, which is understandable.* (Job-seeker, female).

> *I think most of my exams were pretty straightforward but the one I remember most clearly was my GCSE English Language because the exam board had missed half of the poem off the last question so we only got two verses instead of four. Obviously there is nothing you can do about that in an exam. They faxed the rest through and then handed it out twenty minutes before the end which was obviously very stressful for us.* (PGCE student, female).

The participants were knowledgeable about various forms of human error, and if asked to think about the assessment process they could identify where and how errors could occur. However, the participants had not thought of, or labelled such occurrences as 'error'. Indeed, the participants believed that human error occurred so infrequently that it did little to undermine their trust in the reliability of their assessment outcomes.

5.3 AIM 3: ACCEPTABILITY OF RANDOM ERROR AND MEASUREMENT INACCURACY

Measurement inaccuracy was an alien concept to the participants. Although they could identify some causes of measurement inaccuracy such as poor performance on the day, unanticipated questions, or difficult phrasing of questions, they struggled to see how this could (a) be labelled as assessment error or (b) impact on the reliability of outcomes. Instead, the participants tended to see such inaccuracy as an inevitable part of the assessment process, taking a 'that's life' perspective. They suggested that to draw attention to it and label it as assessment error was overkill, particularly as it was unavoidable and could not realistically be eliminated from the process of assessment. Their views tended to mirror some of those expressed by assessment experts attending Ofqual's (2009) Technical Seminar on reliability: that random error is not within the control of the assessors and is therefore less relevant than other forms of error to the issue of reliability. Hayes, for example, (cited in Ofqual, 2009) argues that random measurement error does not lead to valid misclassification, unlike human error, systematic errors, grading errors and clerical errors.

5.3.1 Random error: 'The luck of the draw'

The participants frequently referred to 'luck', or its lack, in relation to their performance in examinations. They tended to oppose the labelling of luck as a form of random error, although they recognised it was an inherent part of examining that could not be controlled. The following quotes demonstrate the participants' awareness of the role that luck played in their examination performances:

> *A lot of it is the luck of which questions turn up on the day, because you can revise and revise and revise and know any subject to death but if the wrong question turns up then it doesn't make any difference how much you know about a subject.* (PGCE student, male).

> *There is a lot of luck in exams. I found that you just have to play it like a game sometimes and people think I'm going to focus on these bits, and if these bits don't come up I'm in trouble.* (Job-seeker, male).

> *Both of the questions I wanted came up, I was absolutely elated, but I don't know what I would have done if they hadn't come up. It just shows you that it's the luck of the draw very often. You can't revise the whole blinking syllabus can you? It's the luck of the draw.* (NHS employer, female).

A recent study of students' perceptions of the provision of stretch and challenge in the new A-levels found that the participants attributed examination success to highly strategic preparation, rather than luck (Baird, Chamberlain, Meadows, Royal-Dawson and Taylor, 2009). The students recognised some element of luck in their successes, but were so familiar with past papers and mark schemes that they believed they could predict with some accuracy the content of their exam papers. Mark schemes were not commonly made available until recently and it may be the case that the influence of 'luck' in explaining an individual's performance has, or is perceived to have, diminished with the increasing use of past papers and mark schemes as learning or revision tools. It is worth noting however, that the participants of the Baird *et al* (2009) study were high-achieving, grade A students. As a delegate at the Ofqual Technical seminar on reliability pointed out, it is students with partial knowledge (as opposed to students who know very little, or a great deal) who are most exposed to 'the luck of the draw' i.e. the randomness of the questions that appear in an exam paper (Ofqual, 2009).

5.3.2 Random error: 'That's life'

The participants appeared relatively ambivalent about random error and measurement inaccuracy. They accepted the occurrence, and could give examples, of such error, but believed it was inappropriate to label it as 'error'. As no one could be held accountable for such errors, the participants preferred to dismiss random errors as being significant factors in determining assessment outcomes. They acknowledged that in some cases they could have gained a different outcome with a different test on a different day. However, they believed that any difference in outcome would be small and would affect scores rather than grades. The participants were somewhat fatalistic about this type of error as the following quotes suggest:

> *On one GCSE I just missed out on a grade C. I was so close, but at the end of the day I'd just missed out on it. Had I worked a bit harder or done a bit more then maybe I wouldn't have and it's just sometimes you win, sometimes you lose and unfortunately that's life isn't it?* (NHS employee, female).

> *I mean* [the assessment system] *tries to be a fair system and no system will be completely fair. I think that's a given and life is sometimes unfair.* (NHS employee, male).

> *It's hard to say but it is a fact of life that it's not fair by the nature of it. Good things happen to bad people, bad things happen to good people. Exams are a part of life.* (NHS employee, male).

> *It's the same with everything in life; sometimes decisions will go in your favour, sometimes they'll go against you. Maybe you'll get marked up on something, maybe you'll get marked down on something else.* (NHS employer, male).

> *There has to be some process by which people are selected. It is a competition for a job or whatever. Not everybody will pass their driving test and that's right, because some people will be better drivers than others. Some people will be great drivers, but not pass their test because of their nerves or the road conditions or something. Life's like that, isn't it? It's not a free-for-all.* (NHS employer, female).

> *Sometimes you're not always going to get absolutely the best deal in the world. You have to live with it. Life's full of poor hands being dealt isn't it?* (Primary school teacher, female).

For some participants, the discussions about measurement inaccuracy appeared to trigger doubt in their minds and led them to bigger questions about the validity of the examinations process. Observing the change in attitude within some participants highlighted the fact that many were considering assessment issues for the first time. It also underlined the importance of having some understanding of measurement error in order to gauge perceptions of reliability. This may be relevant for the Strand 3 questionnaire survey as it suggests that potential respondents may be unable, or unlikely, to respond the survey without a clear understanding of how and why reliability is relevant to their assessment experiences.

> *It's interesting though because, how can an exam accurately pinpoint where someone is in their ability or knowledge? Someone's ability or knowledge isn't a fixed thing anyway so even though a grade is a bit vague and even a specific mark is a bit vague, someone's actual subject knowledge or understanding is a fluid thing. It's not a fixed thing anyway.* (PGCE student, male).

> *Sometimes I think the exams aren't really a fair gauge to a person's ability, are they? They're more about how that person was at the time and how studious they were and all the rest of it. Somebody who's not very good at exams could excel later on and become top of their field.* (NHS employee, male).

Like some assessment experts (Ofqual, 2009), the participants appeared to draw a clear line between human error and measurement inaccuracy. They perceived human error as unfortunate and avoidable, and therefore relevant to the issue of the reliability of outcomes (albeit with minimal impact). In contrast, measurement inaccuracy was viewed as largely irrelevant, being a product of a collection of fortunate or unfortunate circumstances that were no more or less influential in assessment than in other aspects of life. Speakers at the Ofqual (2009) Technical Seminar on reliability share this view and have called for greater clarity in debates about reliability by clearly separating of the forms of error that awarding bodies can and cannot not address. As the following section suggests, however, some of the participants thought the public could be better informed about the assessment process and how measurement inaccuracy could occur.

5.4 AIM 4: REPORTING RELIABILITY

It proved to be crucial that the discussions about human error and measurement inaccuracy took place before the discussion about publishing reliability information. The participants

needed to have a clear understanding about how their assessment outcomes could be affected by error before they could articulate a need, or otherwise, to have access to reliability information. It was evident that many participants were considering reliability for the first time and not conveying their opinions but actually forming – and changing - them throughout the course of the focus group discussion.

The participants offered three perspectives on whether information about the reliability of outcomes should be routinely reported or made available to students, parents, and the public. The number of participants supporting each view was not calculated[5], but the majority of participants were split between the first two views, with a smaller number (mainly the primary school teachers) supporting the third.

1) Some participants believed that the inevitability of measurement inaccuracy, and the fact that nothing can be done to control or prevent it, rendered reliability information redundant and pointless. If nothing could be done to improve reliability over and above current levels (e.g. by eliminating random error), then reliability information would serve no purpose other than to undermine public confidence in assessment outcomes.

> *You know all that anyway don't you, you know instinctively that it's a matter of opinion and subjectivity a lot of it; you don't need to be told* [about reliability] *I don't think.* (NHS employee, male).

> *I think telling people at the end when they get their results back isn't a good idea because that's kind of saying, oh here, look at your results now.* (NHS employee, female).

2) Some participants believed that, if educational experts consider reliability to be an issue (i.e. if it is not the case that grades are 'facts' but rather 'approximations'), then the public should be better informed about the existence of error and its influence of the reliability of outcomes.

> *I suppose if there is a system that isn't 100 percent reliable, that affects the lives of many people, it's not right that people believe it to be 100 percent efficient, so I suppose everyone should be furnished with the facts and do what they will with them.* (Job-seeker, female).

> *I think it would be worthwhile for employers* [to know about assessment reliability]*, and people that have recently gone through the process, people that have just done their A-levels and what have you. I think they all want to know. I'm not sure to what extent the general public wants to know if I'm being honest but I think there is a section that would want to know, employers definitely.* (Job-seeker, male).

> *Education is very important in everybody's life so I think that kind of information should be open and available to everyone.* (NHS employer, female).

3) A smaller number of participants believed that, irrespective of the degree of error in the process, reporting reliability would only undermine the achievements of candidates and create a great deal of uncertainty. They questioned how grades could continue to be used as the

[5] It would be misleading to attribute numbers to each view, as in a focus group a participant might simply nod their head to indicate agreement – a behaviour that would not be evident from the transcript.

basis of selection for education, employment or training if society, and especially employers and universities, did not trust in their reliability. Whatever the benefits to 'society', the participants believed that these could not outweigh the negative impacts on candidates.

> *I wonder if the students were made privy to the information and it was delivered very directly, you know, well: 'so many percent of test results are wrong'. I wonder how it would affect confidence and motivation amongst them?* (Primary school teacher).

> *For an 18 year old, to get a result like that and you think, well it might not be right, I don't think it can do them any good. I don't think it can be good for them. By the time your results go off they need to be results; just results, don't they?* (Primary school teacher, female).

The data collected for this study cannot be applied to determine the extent to which the public wishes to be informed of reliability issues; it was designed to explore only the diversity of opinion. However, as the next section outlines, the secondary school teachers offered some strong pointers as to how reliability information could be best disseminated and would be most usefully received.

5.5 AIM 5: HOW TO REPORT RELIABILITY

Irrespective of whether participants supported the release of reliability information, almost all participants objected to the possibility of reporting a reliability statistic. They suggested that a statistic would be difficult to understand and interpret and thus would be of little use. Reporting a statistic (e.g. confidence intervals or a reliability coefficient) alongside a candidate's grade was seen as especially undesirable, and would lead candidates to question the value of their efforts and the accuracy of their outcome, as the following participants suggest:

> *I think if people want the information it should be freely available; I don't think they should be told every time their child takes an exam… I don't think there should be a slip with the exam results that says, 'Warning; these are only 70% accurate'.* (Job-seeker, female).

> *Given that a statistic will make a slightly fuzzy grade a bit more fuzzier, will it be useful? There's no point in doing these things if the grade that you get at the end of it isn't useful and if you decide that you've got such and such a grade but it doesn't really represent that much and we can't rely too heavily on it then what's the point in doing it to begin with?* (PGCE student, male).

> *Tell them there is uncertainty. Just don't give them a statistic. Say 'there is uncertainty in grades that we cannot escape'. But don't tell them it's thirty percent unreliable, or whatever, because that's a Daily Mirror headline. I think* [Ofqual/the assessment community] *just needs to change people's attitudes towards what a grade is.* (PGCE student, female).

Those who supported the release of reliability information suggested that they simply wanted to be better informed about the process. The secondary school teachers in particular felt

strongly that it is teachers and students who need to be better informed – that the process of educating the public about reliability should start with those at the heart of the assessment system. The following three-person exchange shows the secondary school teachers' lack of support for giving information directly to parents and, instead, their support for being better informed themselves:

Female: "I think the idea of giving [reliability] information to parents would simply muddy the water."
Male 1: "I agree. I'd give that a huge miss."
Male 2: "If anybody was going to spend money informing the parents, I'd rather they spent it organising meetings with teachers where they could sit and talk to examiners."

The teachers felt that if they were better informed they would be in a strong position to convey reliability issues to students and parents and, to some degree, to defend candidates' outcomes. Although the teachers were knowledgeable about reliability and measurement issues, the following quotes suggest that they doubted that they, and teachers generally, fully understood the assessment process:

> *If they could deliver the teacher standardising in a way that helps those teachers understand how the marks are awarded – which is what they do for the examiners – then the teachers would be able to convey more 'I feel confident in the way…It's all good', to the parents. Bypassing the teachers and trying to convince the parents that everything's okay is going to add to the confusion because the parents will just say, 'I disagree'. What has to happen, I think, is that the teachers that are delivering the marks or delivering the syllabus, they have to understand. They don't have to agree but they have to understand and they have to have faith.* (Secondary school teacher, male).

> *When you understand how you apply the standard that actually gives you a bit more confidence in it because you know why you're doing what you're doing. You can tell the kids, 'You need to do this because doing x will give you y at the end of it'. When I talk to parents, I can talk about the way I'm teaching because I know that doing x is going to give them y. Some of it I don't necessarily agree with but I understand it and that's fine. It's the understanding that's key, I think. We don't have to agree with it, we just have to understand it.* (Secondary school teacher, male).

The other participant groups were unable to offer any other perspectives on how to best disseminate reliability information; no doubt reflecting the fact that they had rarely, if ever, considered such issues prior to the focus group discussion. The teachers' view that dissemination should begin with teachers and students is perhaps one of many alternatives. It may be the case that the Strand 3 questionnaire survey will generate support for an alternative way forward.

6 HYPOTHESES FOR THE STRAND 3 QUESTIONNAIRE SURVEY

This section presents the findings as hypotheses, grouped by the five aims of the study. The hypotheses are offered to generate ideas for potential items for the Strand 3 questionnaire survey. The hypotheses reflect the diversity of opinion, and as such are less internally consistent and concise than 'formal' hypotheses.

Public perceptions of reliability.

- The public rarely think about the reliability of assessment outcomes.
- The public trusts that awarding bodies have systems in place to ensure that candidates receive the grades that they deserve.
- The public trusts that examiners are subject experts.
- The public is more concerned about outcomes that are poorer than expected, than about outcomes that are better than expected.

Acceptability of human error.

- The public believes that given the scale of the public examination process, some human error is inevitable.
- The public has a low level of tolerance for typographic errors and errors concerning the distribution of examination materials.

Acceptability of measurement inaccuracy.

- Measurement inaccuracy is caused by a collection of fortunate and unfortunate circumstances that apply to assessment as equally as to other aspects of life.
- The public does not perceive random error as 'true' error as it can not be eliminated from the assessment process.
- The public understands that a different test on a different day may produce a different outcome.
- The public believes that measurement inaccuracy impacts only on assessment scores.
- The public believes that measurement inaccuracy has little impact on a candidate's grade.
- The public does not perceive measurement inaccuracy as relevant to the reliability of assessment outcomes.
- Giving the public examples of random error may encourage qualification-users to view grades as 'approximations' rather than 'facts'.

Reporting reliability.

- The sectors of the population most likely to be interested in and engage with issues of reliability are those who are directly or indirectly involved in educational assessment (e.g. teachers, students, employers, universities).
- Assessment reliability is of limited relevance to the general public.
- Qualification-users should be informed about how errors can occur in the assessment system.
- The public believes that employers should be aware that grades are 'approximations' rather than 'facts'.

How to report reliability.

- The public will not appreciate or engage with the publication of reliability statistics.
- The public believes that the publication of a reliability statistic alongside candidates' grades will confuse candidates.
- The public believes that the publication of a reliability statistic alongside candidates' grades will devalue candidates' achievements.
- Any efforts to improve understanding about reliability should start with teachers and students.

7 CONCLUDING COMMENTS

The aim of this project was to explore public perceptions of assessment reliability and the usefulness of publishing reliability information. In exploring the views of various groups with a direct or indirect interest in educational assessment, this study gives an insight into the diversity of opinion regarding assessment reliability, and provides a measure of the extent to which the public wishes, or feels it necessary, to be more informed about reliability issues. The report has not sought to quantify support for one perspective or another, and instead focuses on representing the views of as many participants as possible.

The focus groups were, for many participants, a first opportunity to engage with reliability issues. Indeed, many participants remarked that they had enjoyed discussing assessment. They suggested that the discussion had given them a new perspective on assessment and reliability and that they felt better equipped to evaluate the usefulness of assessment outcomes, whether as a consumer of educational qualifications or someone involved in the process of selection for employment or education. Throughout the course of the discussion a significant number of participants changed their perspectives - from one of almost blind faith in the reliability of outcomes, to one of 'informed' caution. Any such changes were not researcher-led. Instead, it is attributable to the focus group context, the sharing of experiences, ideas, and opinions, and the participants being encouraged to discuss issues they had not previously had the need or opportunity to think about and articulate. The participants also responded well to the vignettes and found that they were an accessible means by which to engage with issues of reliability. On this basis, Ofqual may wish to consider the inclusion of short case studies or scenarios as part of the programme to communicate reliability to the public or particular qualification-user groups.

On the whole, the participants believed it would be useful to understand assessment reliability and how error can occur. This was mostly in relation to human error (rather than measurement inaccuracy) and focused on students, teachers, employers and universities (rather than parents and the general public). However, it appears that any initiative to enhance understanding of reliability would need to be carefully managed. The participants were typically well educated, but many struggled to comprehend the concept of reliability (even in layperson terms) and to envisage how it was relevant to their experiences of assessment. Similarly, the lack of support for any routine quantification of reliability may be explained by the participants' unfamiliarity with reliability statistics, or statistics in general, and their potential usefulness and application. It may be the case that, in order to engage with reliability issues, the 'public' or a particular target audience may be more receptive to a discursive, qualitative approach to reliability.

8 REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baird, J., Chamberlain, S., Meadows, M., Royal-Dawson, L. and Taylor, R. (2009). *Students' Views of Stretch and Challenge in A-Level Examinations*. London: Ofqual.

Baird, J., Greatorex, J. and Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education*, 11(3), 331-348.

Barter, C. and Renold, E. (1999). *The use of vignettes in qualitative research*. University of Surrey: Social Research Update, Issue 25.

Crisp, V., Sweiry, E., Ahmed, A. and Pollitt, A. (2008). Tales of the expected: the influence of students' expectations on question validity and implications for writing exam questions. *Educational Research,* 50(1), 95-115.

Delap, M.R. (1994). An investigation into the accuracy of A-level predicted grades. *Educational Research*, 36, 135-148.

Delap, M.R. (1995). Teachers' Estimates of Candidates' Performances in Public Examinations. *Assessment in Education*, 2 (1), 75-92.

Denscombe, M. (2000). Social conditions for stress: young people's experience of doing GCSEs. *British Educational Research Journal,* 26 (3), 359-374.

Dhillon, D. (2003). *Teachers' estimates of candidates' grades: Curriculum 2000 Advanced Level Qualifications.* AQA Research Paper RC223.

Flores, J.G. and Alonso, C.G. (1995). Using focus groups in educational research. *Evaluation Review,* 19(1), 84-101.

Gibbs, A. (1999). *Focus groups*. University of Surrey: Social Research Update, Issue 19.

Glasser, T.L. and Salmon, C.T. (1995). *Public opinion and the communication of consent*. New York: The Guilford Press.

Goldstein, H. (2001). Using pupil performance data for judging schools and teachers: scope and limitations. *British Educational Research Journal,* 27 (4), 433-442.

Hammersley, M. and Gomm, R. (1997). Bias in Social Research. *Sociological Research Online,* 2(1). Available from: http://www.socresonline.org.uk/socresonline/2/1/2.html

Ipsos MORI (2009). *Public perceptions of reliability in examinations*. London: Ofqual.

Kitzinger, J. (1994). The methodology of focus groups: The importance of interactions between research participants. *Sociology of Health and Illness,* 16, 103-121.

Krueger, R.A. and Casey, M.A. (2000). *Focus groups: A practical guide for applied research*. London: Sage Publications Ltd.

Linn, R.L. (Ed.) (1993). *Educational Measurement* (3rd ed.). Phoenix, AZ: Oryx Press.

Meadows, M. and Baird, J. (2006). *What is the right mark? Respecting other examiners' views in a community of practice.* AQA Research Paper RPA_06_MM_RP_019.

Meadows, M. and Billington, L. (2005). *A Review of the Literature on Marking Reliability.* AQA Research Paper RPA_05_MM_RP_05.

Morgan, D.L. (1997). *Focus group as qualitative research* (2nd ed.). London: Sage.

MORI (2003). *Trust in public institutions.* London: MORI Social Research Institute.

Murphy, R. (1979). Teachers' Assessments and GCE Results Compared. *Educational Research*, 22, 54-59.

Murphy, R. (2004). *Grades of uncertainty. Reviewing the uses and misuses of examination grades.* London: Association of Teachers and Lecturers.

Newton, P. (2003). The defensibility of national curriculum assessment in England. *Research Papers in Education*, 18(2), 101-127.

Newton, P. (2005). The public understanding of measurement inaccuracy. *British Educational Research Journal*, 31(4), 419-442.

Nunnally, J.C. and Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). London: McGraw Hill, Inc.

Ofqual (2009). *The Reliability Programme: Technical Seminar Report – 7 October 2009.* Coventry: Ofqual.

O'Neill, O. (2002). *A question of trust: The BBC Reith Lectures 2002*. Cambridge: Cambridge University Press.

Pinot de Moira, A. (2005). *Do examiner characteristics affect marking reliability?* AQA Research Paper RPA_05_APM_RP_04.

Please, N.W. (1971) Estimation of the proportion of examination candidates who are wrongly graded. *British Journal of Mathematics, Statistics & Psychology*, v24 p230-238.

Pollitt, A., Ahmed, A., Baird, J., Tognolini, J. and Davidson, M. (2007). *Improving the Quality of GCSE Assessment.* London: QCA.

Ritchie, J. and Spencer, L. (1994). Qualitative data analysis for applied policy research. In Bryman, A. and Burgess, R.G. (Eds.). *Analysing qualitative data*. London: Routledge.

Taylor, R. (2007). *A qualitative exploration of key stakeholders' perceptions and opinions of awarding body marking procedures.* AQA Research Paper RPA_07_RT_RP_034.

Weiner, B. (1992). *Human motivation: Metaphors, theories, and research*. Newbury Park, CA: SAGE Publications.

Wiliam, D. (2003). National Curriculum assessment: How to make it better. *Research Papers in Education*, 18(2), 129-136.

Wiliam, D. (10 October, 2007). *Comparative analysis of assessment practice and progress in the UK and USA.* Presentation at Westminster Education Forum Seminar on Assessment.

APPENDIX 1: VIGNETTES

PGCE students, primary teachers
Teacher assessment
A teacher has provided a set of predicted grades for his class of 28 students, based on homework and their performance in class and class tests. On receiving their GCSE results the teacher sees that he was one or two grades out for approximately half of his student group. How might this have happened? Did the teacher get it wrong or is the exam result wrong?

NHS employees
Reliability of assessment outcomes
The owner of a small business has advertised for a full-time Finance Assistant. She can only conduct a limited number of interviews and is finding it difficult to choose between two school-leaver applicants. The applicants studied the same subjects and received the same grades. One of the applicants completed their exams at the end of their two year course. The other student completed their exams as they covered each topic and completed coursework that was marked by their teacher. One of the applicants went to a prestigious private school, while the other went to a large inner-city comprehensive. Does having the same grades mean that the applicants have the same abilities?

NHS employees, NHS employers
Application of grades
Two students are applying for the same university course which requires A Level grades of AAB. One student achieves the required grades and is accepted on to the course. The other student achieves ABB and is rejected by the university. Why might the student not have achieved the required grades? Is it fair that they were rejected?

NHS employers
Vocational versus academic qualifications
A young mother is diagnosed with a non-terminal illness. As part of her care programme she is given the choice of two newly-qualified specialist medics. One of the medics passed several knowledge-based written examinations. The other medic was mostly assessed on their clinical skills and passed several patient assessments. Does it matter which medic the patient chooses?

Job seekers
Assessment error
Twin brothers were predicted grade B for A-level Chemistry. One of the twins got his grade B, but the other twin was very disappointed that he got a grade D. What kind of things could have gone right for the first brother, and what could have gone wrong for the second brother?

PGCE students, Job seekers
Assessment error
The owner of a small business has advertised for a full-time Finance Assistant. She can only conduct a limited number of interviews and is finding it difficult to choose between two school-leaver applicants. One of the applicants has grades AAB in GCSE Maths, English and Science, and the other applicant has ABB. Would it be fair to say that the AAB applicant will be more knowledgeable than the ABB applicant? Would it be fair if she then put the ABB applicant's form in the bin?

First published by The Office of the Qualifications and Examinations Regulator in 2010.