Marking as judgment

Val Brooks[1]

Institute of Education, University of Warwick, Coventry, UK

An aspect of assessment which has received little attention compared with perennial concerns, such as standards or reliability, is the role of judgment in marking. This paper explores marking as an act of judgment, paying particular attention to the nature of judgment and the processes involved. It brings together studies which have explored marking from a psychological perspective for the purpose of critical discussion of the light they shed on each other and on the practice of marking. Later stages speculate on recent developments in psychology and neuroscience which may cast further light on educational assessment.

Keywords: marking; assessment; judgment

Introduction

Marking is a multi-faceted topic which has been explored from perspectives as diverse as the technical requirements of validity and reliability to its use as a formative tool to enhance learning. An aspect of marking which has received comparatively little attention until recently is the role of judgment. Most assessments – with the exception of those adopting a fixed-response format which enables marking to be completed by a machine – are underpinned by the judgment of individuals. Despite its pivotal role in assessment, the nature of judgment and the processes involved are topics which have received scant attention compared with perennial concerns such as standards or the reliability of marking. Systematic attempts to explore the role of judgment in marking are a recent phenomenon with researchers invariably expressing a sense of venturing into little explored terrain – a point which has been observed in the context of schoolteacher assessment (Wyatt-Smith and Castleton 2005), external examining (Suto and Greatorex 2008) and higher education (Elander 2004). In other disciplines, medicine for instance, research into judgment has played a role in the development of theory and the improvement of practice

[1] Email: v.brooks@warwick.ac.uk

(Chapman 2005)) suggesting that there may be much to gain from closer attention to this aspect of marking.

This paper explores marking as an act of judgment, bringing together studies which have explored marking from a psychological perspective for the purpose of critical discussion of the light they shed on each other and on the practice of marking. Later stages speculate on recent developments in psychology and neuroscience which may also elucidate educational assessment. The following section sets the scene by considering judgment as a research topic.

Research Context

Much of the research into judgment has taken place in disciplines other than educational assessment: business management, economics, accounting, medicine, public policy and governance. Elander (2004, 114) suggests that this 'relative lack of attention' is 'perhaps surprising' given that assessment is 'part of the natural subject matter of psychology'. One consequence of this relative neglect is that: 'Within the very broad field of psychology there exist multiple constructs of judgement and decision-making, which have yet to be applied to examination marking' (Suto and Greatorex 2008, 214). An overarching concern of extant research is to expound the relationship between rational and intuitive thought. The branches of Psychology that have been drawn on most frequently to analyse and interpret markers' behaviour focus on bounded rationality – the shortcuts used to cope with limited time and the limited information processing capacity of the human mind – and judgment under uncertainty – how judgment is affected when the information available is insufficient to provide an adequate basis for judging. Research has focused on: biases, which are sometimes viewed as concomitants of heuristics; aspects (or cues) that subjects attend to as they mark; the role of memory, especially the working memory; tacit knowledge

and the formation of social and individual cognitive constructs. However, researchers who have addressed the topic from an educational perspective have sometimes adopted a grounded approach, allowing an explanation of findings to emerge from the analysis of data with little or no explicit reference to extant theories of judgment (e.g. Morgan 1996). Although various theoretical stances have been adopted, the volume of research remains comparatively slim.

The choice of research methods is a particularly important consideration because judgment is a tacit process which leaves no trace of its workings under normal circumstances. These are but two of the issues which make judgment elusive and notoriously difficult to study. The challenge of elucidating processes which are normally cloaked by the impenetrability of the individual's mind has encouraged the use of more innovative and technically sophisticated methods as well as standard methods such as interviewing. For instance, judgment analysis is used to capture an expert's judgment policy by analysing a number of prior judgments to determine the different cues (or aspects) which are attended to and how these are weighted and combined. This information is used to create a statistical model of the individual's judgment policy which can be applied to future cases involving similar judgments. Elander (2004, 118) notes that this mechanical method can eliminate errors and bias, leading to 'better decisions than those based on an expert's intuitive or holistic judgement'. However, Cooksey, Freebody, and Wyatt-Smith (2007, 428) concluded that teacher judgment involves a process that: 'is more complex than can be represented in psychometric models or in linear judgment models', arguing that combining judgment analysis with think aloud: 'elaborates on a story that neither alone can tell in full' (429). Indeed, it is widely accepted that a combination of methods yields better information than reliance on a single method. For instance, van

Someren, Barnard, and Sandberg (1996, 26) recommend that one method may be used to 'focus or facilitate application of the next'. Eraut (2000, 120) suggests using 'a mediating object like a picture or a drawing' as a stimulus. A study of the clinical knowledge of nurses found that a twin methodology combining interviews with knowledge maps and digital photographs within an hour of observed events afforded a 'unique perspective' on clinical expertise (Fessey 2002, 47).

Think aloud is a commonly used elicitation method. It requires subjects to generate concurrent verbal reports of their thoughts whilst they are marking, making it a useful tool for eliciting cognitive processes which would otherwise remain tacit. Whereas subjects are inclined to manipulate reports that are given retrospectively, a strength of concurrent reporting is that: 'Because almost all of the subject's conscious effort is aimed at [the task in hand], there is no room left for reflecting on what he or she is doing …He or she renders [thoughts] just as they come to mind' (Van Someren, Barnard, and Sandberg 1996, 25-26). However, think aloud is not without difficulties. For instance, Ericsson (2002) stresses that the instructions and procedures used strongly influence the capacity of think aloud to generate faithful verbatim accounts of thoughts. He calls for greater methodological rigour and standardisation of approaches, along with explicit accounts of these aspects of research. Van Someren, Barnard, and Sandberg (1996, 34) also note that experts are inclined to be 'secretive' or 'reluctant to give someone else insight in their actual problem-solving behaviour', behaving 'more rationally' than they would in natural settings. Furthermore, Leighton (2004, 11) has cautioned against treating think aloud as a tool which is equally adapted to all situations where one wishes to elicit thought. Difficult tasks tend to yield sparse verbal reports because they 'take up a lot of mental resources and overload working memory' leaving 'few, if any, working memory resources to

actually describe or articulate the process'. Easy tasks are also ill-adapted to verbal reporting because they rely on automatic cognitive processing which 'occurs too quickly failing to leave a conscious trace in working memory' (11). Thus, think aloud works best with tasks of 'moderate difficulty' (11). Finally, influences which operate below the cognitive threshold are unlikely to be captured by think aloud as subjects will not verbalise that of which they are unaware. Experimental research plays an important role in uncovering aspects of judgment which are pre-conscious or which occur too rapidly to be perceived (e.g. Laming 2004).

Overall, the obstacles to elucidating judgment remain considerable. Whilst acknowledging that researchers should 'reach as far as they can down the continuum from explicit to tacit knowledge' (Eraut 2000, 119), Eraut urges caution, noting that: 'the limitations in making tacit knowledge explicit are formidable … There can be many benefits from making some progress in this area … Nevertheless researchers need to be both inventive and modest with their aspirations' (135).

Judgment processes in marking: Are they qualitatively different?

Keren and Teigen (2004, 93) characterise the psychology of judgment thus: 'It may be slow and deliberate, like problem-solving, and quick and immediate, like for instance distance perception, where we seemingly jump to the conclusion (e.g. "a car is approaching")'. The relationship between rational and intuitive thought has been an important preoccupation for recent assessment research. For instance, Suto and Greatorex (2008) used a recent development in the heuristics and biases research programme (see section, 'The role of heuristics?') – the dual processing theory of judgment – to explain the behaviour of General Certificate of Secondary Education (GCSE) examiners. This theory distinguishes 'quick and associative' system 1 judgements from 'slow and rule-governed' System 2 judgements (Suto and Greatorex

2008, 215). Judgments made using System 1 are 'intuitive', 'automatic, effortless, skilled actions, comprising opaque thought processes, which occur in parallel and so rapidly that they can be difficult to elucidate' whereas System 2 judgments involve 'slow, serial, controlled and effortful rule applications, of which the thinker is self-aware' (215). The two systems are thought to be concurrently active, enabling subjects to switch between them according to the cognitive demands of the task in hand. Suto and Greatorex (2008) used this theory to interpret the cognitive strategies employed by GCSE examiners assessing two subjects chosen for their contrasting contents and mark schemes: a points-based mark scheme in mathematics and a primarily levels-based scheme in business studies. Six examiners in each subject marked scripts from the previous year's examinations. Suto and Greatorex combined think aloud with semi-structured interviews, identifying five distinct cognitive marking strategies. For instance, the 'matching' strategy required 'a simple judgment of whether a candidate's response matches the mark scheme' (220). It was presented as system 1 judgment because markers could rely on rapid pattern recognition, identifying, for instance, a word, letter or number which matched the mark scheme. 'Scrutinising' (225), in contrast, was used for unexpected responses where a marker needed to determine whether an answer was due to error or represented an acceptable alternative to the mark scheme. Scrutinising was presented as evidence of system 2 judgment because it entailed multiple rereads of a text, pauses, hesitations and recourse to the mark scheme as markers tried to resolve their uncertainty. Another important feature of the dual processing theory is that 'complex cognitive operations may migrate from System 2 to System 1' (215) as individuals gain experience. Indeed, some 'very experienced examiners', who Suto and Greatorex consulted about

their findings, raised concerns that some assistant examiners were switching from System 2 to System 1 on particular questions 'before they were ready to do so' (229).

Ecclestone (2001) is another proponent of the view that the judgments which underpin assessment are qualitatively different. Her study entailed some important differences from that undertaken by Suto and Greatorex (2008) yet her conclusions bear some striking similarities. First, Ecclestone investigated the marking of undergraduate dissertations whereas Suto and Greatorex focused on marking at GCSE level. Thus, Ecclestone's markers assessed longer, more complex responses than those judged by the GCSE examiners. The dissertations were completed as part of an Education degree whereas Suto and Greatorex used scripts from Mathematics and Business Studies examinations. Ecclestone also adopted different research methods: a two-year case study during which she acted as a participant observer in annual moderation meetings as well as interviewing markers, analysing their written feedback and comparing grades awarded. Finally, Ecclestone drew on a different theoretical source, Eraut (1996) who, in turn, drew on Dreyfus and Dreyfus' model of professional decision-making. The resulting model of judgment is more elaborate than the dichotomous system proposed by Suto and Greatorex in that it distinguishes four categories, each linked to a stage in the development of expertise: novices, advanced beginners, competents and experts. Arguably the most important difference between these theories is that the model adopted by Ecclestone is expertise-based and follows a step-wise approach, with earlier stages being superseded by subsequent stages, whereas in dual processing the two systems remain concurrently active, enabling assessors to switch between them according to the demands of the task. Ecclestone's model does not preclude the use by experts of approaches associated with earlier stages. On the contrary, this is recommended as a means of countering the 'erratic

interpretation' to which expert judgment is prone (305). However, this is presented as problematical because experts are resistant to using approaches associated with inexperienced status. Despite the considerable differences between these studies, Ecclestone's characterisation of expert judgement has much in common with Suto and Greatorex's account of system 1 judgment. For instance, expert judgment is depicted as displaying 'a declining dependence on rules, routines and explicit deliberation. Experts become more intuitive and less deliberative, and are less able to articulate the tacit knowledge on which much of their decision making has come to depend' (305). Likewise, novices were found to be more reliant on rule-based guidelines making novice judgment akin to system 2 judgment. Both studies suggested that the accumulation of experience inclined markers to speedier, less considered judgments.

The points on which these theories concur may elucidate findings reported elsewhere which might otherwise appear baffling. For instance, an American study compared the scores awarded by five middle school mathematics teachers, who had been trained to use a rubric for marking non-traditional mathematical tasks, with the scores given by an expert group composed of mathematics education researchers who used the same rubric (Meier, Rich, and Cady 2006). The results indicated that four of the five teachers experienced difficulty when required to give equal consideration to two separate criteria (the correctness of procedures and fullness of explanations), focusing on one at the expense of the other. When the task was less complex (i.e. only one criterion was involved), 'teachers were better able to judge the student responses using the rubric' (91). However, they experienced greatest difficulty in giving equal weight to two separate criteria when the mathematical task involved familiar content. When the content was unfamiliar, the level of difficulty was reduced. On the face of it, this is counter-intuitive. One would expect unfamiliar content to compound the

level of difficulty by increasing the cognitive load, yet the opposite appeared to be true. Meier, Rich, and Cady speculate that: 'perhaps … when the content is unfamiliar the teachers must think more about the task, and the processes and explanations themselves, making both important. The familiar tasks do not require the teachers to think about the reasoning: thus, they do not look carefully at the explanations' (90-91). This proposition articulates with the theories outlined above, suggesting that an element of unfamiliarity in the content may have obliged these markers to make qualitatively different judgments, behaving less like 'experts', making 'system 1' judgments, and more like 'novices' using 'system 2' thought processes. This led to improved construct representation (correctness of procedures and fullness of explanations) enhancing the construct validity of the assessment.

It is important to note that this is contested territory and that the theory of qualitatively different judgments is not universally accepted within the discipline of Psychology. An alternative view holds that 'so-called System 2 judgment is actually a collection of closely linked System 1-type judgements, occurring both in parallel and in series … From this perspective, the evaluating strategy would comprise multiple rapidly occurring judgements, of which only the combined results are being verbalised' (Suto and Greatorex 2008, 224-225). If this stance is accepted, it also casts doubt upon the notion that judges pass through stages, moving from a slower, more controlled type of judgement to a more rapid and intuitive mode as they acquire expertise. Alternatively, it can be argued that confidence in the claim that there are qualitative differences in the nature of the judgments underpinning assessment is increased by the fact that different studies, involving a variety of contexts and assessment tasks, have yielded similar characterisations of the underlying judgment processes. Another alternative, which emerges later in this paper, is that an either/or

stance simplifies what is, almost always, a complex amalgam combining elements of rational and intuitive thought.

Comparators for judgment: Published criteria

Judgment does not take place in a vacuum; it requires some form of comparator. In assessment, these comparators are known as 'referents' and a number of these are in common use: criteria, constructs, self (or ipsative assessment) and norms (Wiliam 1992, 17). The theory that there is a type of judgment which is algorithmic and deliberative is compatible with an approach to assessment which uses published criteria as referents. Published criteria are widely regarded as one of the principal means of enhancing the transparency, consistency and fairness of assessment and there is an extensive literature critiquing their capacity to deliver these goals (e.g. Jonsson and Svingby 2007; Price and Rust 1999). However, an international survey found that even within a single sector, higher education, 'there is no common understanding of what criteria-based means or what it implies for practice … Additionally, the concepts of "criteria" and "standards" are often confused' (Sadler 2005, 175).This paper adopts Sadler's definition of a criterion as 'A distinguishing property or characteristic of any thing, by which its quality can be judged or estimated, or by which a decision or classification can be made … Criteria are attributes or rules that are useful as levers for making judgments' (178-79).

Across educational sectors, it is regarded as good practice for qualifications to publish the criteria they use to assess performance. The criteria, therefore, are pre-determined, often by a panel of experts convened expressly for the purpose. Familiarisation with the criteria becomes a prerequisite for judgment and markers may be required to undergo a period of training during which the meaning of the criteria, and how to apply them, is clarified. Thereafter, they remain central to

assessment, acting as a touchstone for judgment. Yet questions remain about the extent to which this approach articulates with ways in which judgments are actually enacted. For instance, this approach does not fit well with the theory of system 1 judgment nor the theory that experts are inclined to judge in ways which are intuitive and increasingly independent of rules and routines. There is, however, a better fit with theories of novice and system 2 judgments which involve rule application. Empirical research provides evidence supporting the claim that published criteria are important in the induction of novice assessors (Ecclestone 2001; Wolf 1995). It also confirms that experienced assessors are less likely to judge in the way that assessment criteria require. Wolf (1995, 71), for instance, describes a UK study of vocational assessment involving invoice completion. Industry informants had 'insisted that the criterion for competent performance was 100 per cent accuracy: mistakes might be tolerated in school but not in the workplace'. Yet when: 'Asked to assess the invoices against the relevant standards, the experienced and inexperienced behaved totally differently. The inexperienced failed everyone because everyone had failed to meet the criterion: the experienced judged many competent (as had their own workplace supervisors)' (71). Wolf explains this contradictory state of affairs thus: 'All the research evidence that we have on assessors' behaviour emphasizes the very active role that their own concepts and interpretations play … Assessors do not simply "match" candidates' behaviour to assessment instructions in a mechanistic fashion' (67-68).

Similar conclusions were reached by Hay and Macdonald (2008) who used semi-structured interviews and participant observation over a twenty week period to explore the way in which two Australian teachers, responsible for school-based assessment of PE, used the criteria and standards published in the official PE syllabus. Both teachers were found to assess intuitively, relying on their memories of how

pupils had performed during the course and consulting the official criteria retrospectively to seek confirmation of judgments that had already been made [‘“I think that most of us make our judgements first and then we look at the criteria sheet and see if it backs us up”’ one teacher observed (160)]. The teachers justified their approach by claiming that they had already internalised the official criteria and standards. However, Hay and Macdonald contend that this process of internalisation resulted in ‘a new set of criteria and standards that bear some semblance to the official set outlined in the syllabus’ (165) but that these were combined with teachers’ idiosyncratic values, beliefs and expectations about performance. One teacher, for instance, valued being ‘switched on’ and enthusiastic, being prepared to contribute to lessons and ask questions. The other assessed his students by taking them on in a game [‘“I do my best and I expect them to meet the challenge”’ (161)]. For this teacher, valued dispositions included a readiness to take him on, a drive to win and aggression in play. Thus, each set of internalised criteria developed by a teacher incorporated ‘construct-irrelevant affective characteristics of students’ (153) that were absent from the official criteria. Hay and Macdonald concluded that this construct irrelevance ‘compromised the construct validity and possible inter-rater reliability of the decisions made and advantaged some students and marginalised others on the basis of characteristics that were not specifically related to the learning expected from following the syllabus’ (153).

Comparators for judgment: individual and social constructs

There is other evidence which questions the use of published criteria as assessment referents. For instance, the finding that communities of markers are capable of developing ‘a general construct of “level” or “ability” … without access to any stated criteria’ (Morgan 1996, 356) would be puzzling if explicit criteria were a necessary

component of a referencing system. Likewise, it would be hard to explain the paradoxical finding that assessors working on modules with tightly defined specifications arrived at very different results yet 'apparently very vague and "woolly" communication modules … turned out to be highly consistent across the group of colleges studied' (Wolf 1995, 77). Findings such as these raise questions about the role of published criteria in the processes of (i) forming judgments (ii) achieving and maintaining consistency between markers. Wiliam (1996) contends that the ability of markers to agree in the absence of criteria is because a different process is at work – construct-referencing – claiming that 'most sophisticated assessments that take place in the real world are of this kind' (Wilam 1992, 19). Jackson (2002, 3) describes constructs as 'people's constructions of important entities in their psychological world'. Amongst the 'important entities' in the 'psychological world' of markers of graded assessments are constructs of the nature of performance which typifies different grades – a phenomenon alluded to by Ecclestone's (2001) title, 'I know a 2:1 when I see it'. It entails recognising the standards embodied in individual performances. Indeed, Wiliam (1992, 19) has cautioned against succumbing to the pressure to criterion-reference all assessments, especially complex skills and performances which are irreducible and cannot be itemised because 'the whole is greater than the sum of parts'.

Empirical evidence supporting this assertion comes from a number of studies involving complex, holistic assessments, although it is often anecdotal. For instance, Wiliam (1996, 297) notes that teachers involved in a GCSE English qualification assessed entirely by coursework 'quite quickly internalized notions of "levelness", so that in the vast majority of cases different teachers would agree that a particular portfolio of work merited, say, a D. This was even though the teachers had no explicit

criteria and did not agree on what aspects of the work were most significant in making the award'. Further evidence appears in an account of the introduction of a new GCSE English syllabus, devised to reflect changes in the National Curriculum for England and Wales. The previous syllabus had supported construct-referenced assessment, allowing teachers to think in terms of grades, which were subsequently converted into marks. However, the new syllabus, 'was by far the most exacting in the way it demanded certain criteria should be met' (Marshall 2000, 162), thereby altering the way in which judgments were to be made. Marshall recounts an incident at a consortium meeting which illustrates the dilemma this new syllabus had occasioned. Discussion focused on a script which, according to the convenor, '"screams D"'. However, the script did not meet all of the new criteria for a D. The convenor, who was there to 'guide teachers through the whole process' of implementing the new syllabus, urged them not to: '"get bogged down in looking at the assessment criteria"' (163) thereby favouring the established construct-referenced approach over the new criterion-referenced system. The incident illustrates the destabilising effect of undermining established constructs, throwing into doubt teachers' confidence in their own judgment. As one examiner remarked: '"I had an understanding of what a D was. I've marked scripts. But I simply don't see it …"' (164).

Whilst personal constructs enable individuals to make sense of their world, there is also evidence that where there are opportunities to collaborate, the development of constructs can become a shared undertaking, emerging within a community of practice (Wenger 1998). This appears to underlie Wolf's (1995) seemingly contradictory finding that assessors working with tightly defined specifications were highly inconsistent whereas those marking vague communication modules achieved high levels of agreement. The explanation of this seeming paradox

was that tutors on the communication courses, concerned by 'their own uncertainty about how to interpret the criteria … had formed a close network to share ideas and interpretations and so developed common understandings' (77). The value of communities of practice for developing shared constructs and enhancing consistency of assessment is a recurring theme in assessment literature. For instance, Wilmut, Wood, and Murphy (1996, 20) noted that 'greater consistency of marking can be achieved when markers work in teams (a "conference" setting) than when they mark singly, even when monitored'. They also noted how the reliability of National Vocational Qualification assessment was enhanced by 'internal verification meetings with assessors' because these meetings promoted 'more consistent assessment practices across assessors, particularly when focused on the interpretation of standards and the sufficiency of evidence' (11). This message has emerged across educational sectors: in vocational assessment, in higher education and in the schools sector where it has been observed that: 'the constructs of "level" of secondary and of primary teachers … are determined more by group membership than by any "objective" meaning that might be attached to the criteria' (Morgan 1996, 356). This is because, despite the '"objective" appearance of the use of an assessment scheme based on either generic criteria or task-specific performance indicators, the practical implementation of the scheme relies on the existence of socially constructed consensus among the assessors about their application' (356). Despite their widely reported benefits, studies from across educational sectors suggest that the time required to develop and sustain assessment communities makes them vulnerable to competition for resources from new initiatives. The influx of new initiatives may divert energies, previously devoted to assessment, causing established communities to atrophy (Garry, McCool, and O'Neill 2005; Hall and Harding 2002).

Judgment under uncertainty: The role of biases

Laming (2004) draws on numerous laboratory experiments to evince the claim that a key feature of judgment under uncertainty is its susceptibility to extraneous influences – in other words, the tendency to become biased. Evans (1993, 16) describes bias as 'systematic attention to some logically irrelevant features of the task, or systematic neglect of a relevant feature'. Laming argues that bias is 'irresistible' (153) because it is pre-conscious, coming into operation whenever the available evidence is insufficient to support judgment: 'To the extent that judgment is uncertain, past experience enters like air rushing in to fill a vacuum' (164). He argues that it is because we all have 'different accumulations of past experience' that we tend to 'make different judgments about the same issue' (18).

This theory may help to explain various aspects of marking including the variable levels of inter-rater reliability reported in marking experiments. Some assessments may be viewed as having a substantial element of uncertainty built into them either because of the methods used (e.g. essays marked using qualitative criteria) and/or because of the inherent nature of the subject (e.g. subjects which prioritise creativity or a personal response such as Art or English). Other assessments may be viewed as less uncertain either because of the methods used (e.g. short, structured items marked using a points-based mark scheme) and/or because of the inherent nature of the subject (e.g. subjects which prioritise mastery of a body of knowledge such as Science). An investigation by Murphy (1978) into the reliability of eight subjects at General Certificate of Education Ordinary and Advanced Level (O-Level and A-Level) is relevant here. The study identified three main influences on reliability: subject area; question type and the number of parts that contributed to a final mark. In connection with subject area, Murphy found that although all of the

examinations contained large proportions of essay questions, the English papers produced the poorest reliability scores. One of the A-level Literature papers and the essay paper in the English Language O-Level yielded levels of reliability markedly lower than those produced by any of the other thirteen papers used in the investigation. The effect of question type was found to be most pronounced in an examination employing exclusively essay-type items. If the findings related to judgment under uncertainty apply to marking in the same way that they have been used to explain other types of decision-making behaviour, they may suggest that the more sources of uncertainty there are in an assessment, the more susceptible to bias that assessment is likely to become, thereby increasing the threats to inter-rater reliability.

Lower levels of inter-rater reliability have indeed been found in examinations where there is a substantial element of uncertainty inherent in the subject and/or assessment method (Newton 1996; Wilmut, Wood, and Murphy 1996). Moreover, a wide range of biases has been detected in educational assessment. The range is too extensive to enumerate here so what follows is intended as an illustration rather than a comprehensive account. First, there is evidence that knowing a student whose work is being assessed allows a range of personal considerations to influence assessment. Students' work habits, social behaviour, gender, ethnicity, cultural background and physical attractiveness are amongst the factors which have been found to bias the assessment of academic performance when students are known by their assessors (e.g. Dennis, Newstead, and Wright 1996; Gipps and Murphy 1994; Harlen 2005; Meier, Rich, and Cady 2006; Pollitt and Murray 1996). It has also been shown that markers are susceptible to the 'halo effect' of finding what they are predisposed to find based on prior knowledge of a student (Harlen 2004). Dennis, Newstead, and Wright (1996,

516) suggest that this is consistent with psychological theories of impression formation: 'individuals tend to form consistent impressions of others at an early stage in the impression formation process … and having done this are prone to discount evidence which is inconsistent with those early views'. In assessment research, impressions have been shown to form on the basis of brief encounters (Pollitt and Murray 1996).

Personal knowledge is not the only source of bias. In situations where students are unknown, such as external examinations, Dennis, Newstead, and Wright (1996) suggest that group stereotypes are a more likely source of bias – for instance, if social group can be deduced from a candidate's name or that of their examination centre. When written performance is assessed, surface features including the neatness and legibility of handwriting and the font size used to word process assignments have been identified as sources of bias (Vaughan 1992; Milanovic, Saville, and Shuhong 1996; Hartley et al. 2006).

Although many studies have focused on the detection of biases, fewer have investigated their operation. A three-year Australian study is pertinent here. Wyatt-Smith and Castleton (2005) used think aloud to investigate how teachers judged the written English of ten-year-olds in three different contexts: in-context judgments involved assessing work by their own students; out-of-context judgments involved marking work samples by unknown ten-year-olds and system context judgments involved re-assessing the anonymous samples against Australian literacy benchmarks. They identified a set of 'indexes' which were used to make in-context judgments, showing how difficult judgment became when a shift to out-of-context assessment rendered certain indexes unavailable. Although these indexes were not presented as biases, the infrequency with which reference was made to official documents and the

national benchmarks was noted. Moreover, some of the indexes may be regarded as consistent with the definition of biases given above, for instance, 'assumed or actual knowledge of the community context in which the school is located' (136). Wyatt-Smith and Castleton found a complex and fluid situation where indexes formed unstable, interactive networks. Thus, the weight attributed to different indexes, and how they were combined, varied 'not only from teacher-to-teacher but also from judgement-to-judgement' (144). They concluded that: 'There is no simple, linear course that teachers follow to arrive at their judgements. On the contrary, what emerges is a picture of how dynamically networked indexes come into (and out of) play in acts of judgement' (135). These conclusions are consistent with findings from research using judgment analysis. For instance, Elander and Hardman (2002, 318) reported that 'judgment policies varied from marker to marker' amongst the seven university lecturers in their study whilst Cooksey, Freebody and Wyatt-Smith (2007) identified forty different judgment models operating amongst twenty primary school teachers.

The role of heuristics?

Biases are viewed as 'markers or signatures' for 'underlying heuristics' (Gilovich and Griffin 2002, 3) by psychologists in the heuristics and biases research tradition. Gilovich and Griffin describe heuristics as 'highly efficient mental shortcuts' (4) which offer simpler and quicker ways of judging than the extensive algorithmic processing that is characteristic of rational thought. They claim that heuristics not only reduce cognitive complexity to dimensions which are more commensurate with the limited capacity of human judgment but that they represent a more natural mode of thought than the application of reason. A range of general purpose and specialised heuristics has been identified, each accompanied by associated biases. Although the

detection of biases has been a preoccupation of assessment research, there has been little work exploring the potential of heuristics as underlying, explanatory factors.

One study which used heuristics to explain the biases evident in markers' behaviour was designed to test the hypothesis that the difference between a first marker's grades and a moderator's grades would be less than that between two 'blind' markers' grades (Garry, McCool, and O'Neill 2005). The biasing effect of knowing a previous assessor's marks has been known about for many years (e.g. Murphy 1979). Garry, McCool, and O'Neill illustrated its operation in a higher education setting, using twenty two Politics lecturers to undertake double blind marking of eleven undergraduate examination answers. A second phase of the study entailed the re-distribution of marked scripts amongst participating lecturers who then moderated the initial marking. As predicted, the differences between marks awarded blind were 'much greater in size' (194). However, the distinctive contribution of Garry, McCool, and O'Neill entailed using the anchoring and adjustment heuristic to explain this observation. This heuristic draws on the finding that people sometimes reduce uncertainty by starting with an 'anchor' or 'beginning reference point' (191) which is adjusted to reach a final conclusion. The anchoring and adjustment heuristic has been observed in various settings, mock jury trials for instance. Thus, when half the jurors in a mock jury trial were instructed by the judge to start their deliberations by considering the harshest verdict possible whilst the other half were instructed to start by considering the most lenient sentence possible, the first jury delivered a much harsher verdict than the second. This is consistent with the theory that the judge's instructions had acted as an anchor which was adjusted to reach a final verdict. Garry, McCool, and O'Neill offer a similar explanation of their finding that greater

discrepancies arose when marking was conducted double blind than when first marking was moderated.

The extent to which markers use heuristics, which heuristics are prevalent or whether there are any special purpose assessment heuristics, are topics which have been little explored. Therefore, what follows is necessarily speculative. It suggests that heuristics could provide plausible explanations for various behaviours that have been observed during studies of marking. For instance, a key feature of heuristics is that they make cognitive tasks more manageable by reducing complexity. One area where this may have relevance is in explaining markers' use of published assessment criteria. Markers have repeatedly been found to reduce and/or simplify criteria, a practice which impacts on the construct validity of an assessment. For instance, Bridges, Elliott, and McKee (1995, 6) reported that teacher educators who attempted to apply the criteria devised by the Department for Education to assess student teachers experienced difficulty in applying the full range: 'In practice we found that teacher educators tended to reduce the lists of competences specified by the Department for Education to not more than six broad categories'. Likewise, although the nine tutors in Vaughan's study completed their marking using eight assessment criteria, five different reading strategies were identified. Two were characterised by a 'single-focus', whilst a third was described as 'the "two-category" strategy' and the fourth as 'the "first impression dominates"' (Vaughan 1992, 118). Further evidence appears in a study of Key Stage 3 National Curriculum assessment in English which reported that: 'markers failed to make distinctions between the mechanics of writing and the candidates' capacity to demonstrate understanding and write expressively; they generally failed to reward the latter' (Wilmut, Wood, and Murphy 1996, 20). Wood (in Wilmut, Wood, and Murphy 1996, 20) also found that markers were '"quite

unable to distinguish (at least when marking) between different features of writing"'. Findings such as these are consistent with Tversky and Kahneman's (2002, 20) claim that: 'One of the manifestations of a heuristic is the relative neglect of other considerations'.

Working memory and heuristics

Investigations into the functioning of the working memory (e.g. Baddeley 1998) are also pertinent here. Grimley, Dahraei, and Riding (2008, 214) define working memory as 'the temporary storage of information that is necessary for performing cognitive tasks', emphasising that: 'A practical feature of working memory is its limited capacity and the vulnerability to loss of information in it from displacement by further incoming information'. A study by Scharaschkin and Baird illustrates how working memory shortages have been used to explain judgment in assessment. Scharaschkin and Baird (2000, 343) investigated the role of expert judgment by examiners in the setting of A-Level standards in Biology and Sociology, focusing on the puzzling observation that 'there is not a one-to-one correspondence between marks and grades'. In other words, examiners may consider two scripts to be worthy of the same overall mark but of different grades. They investigated three conditions – consistent, average and inconsistent performance – based on the range of marks awarded to a script and focusing on the A/B and E/N grade borderlines in each subject. Participants were asked to award each script a grade and to rate their difficulty in awarding a grade on a five-point scale. The pattern of results differed in each subject. In Biology, inconsistent performance produced lower judgments of grade-worthiness than average or consistent performance whereas in Sociology there was a preference for very consistent performance. Overall, consistency of performance emerged as a statistically significant factor in judgments about grade

worthiness. Scharaschkin and Baird attributed their findings to the difficulty of multivariate decision-making tasks which require the integration of many different pieces of information, pointing to research which found that people could not hold two distinct dimensions in mind. They argued that 'Lack of consistency of performance probably affects judgements because it forces examiners to integrate contradictory information about a candidate's performance' (354). Therefore, because 'people are poor at holding different states of the world in mind due to working memory constraints … It is likely that examiners' grading judgements are erroneously affected by consistency of performance' (354).

Although Scharaschkin and Baird did not include heuristics and biases in their interpretation, they may offer a perspective on this phenomenon. According to the definition of a bias given above, these A-Level examiners were making biased judgments because consistency 'was not part of the marking scheme' (343). Indeed, Scharaschkin and Baird argued that 'Examiners would probably not wish consistency of performance to be taken into account in a marking scheme' (354). It is possible, therefore, that this bias was the 'marker' or 'signature' for an underlying heuristic. Representativeness is a general purpose heuristics which is based on mental models, such as prototypes, and provides an assessment of the degree of correspondence between, for instance, a sample and its parent population or an instance and a category (Tversky and Kahneman 2002). Although the pattern of results differed in Sociology and Biology, the representativeness heuristic may help to explain why consistent performance was a statistically significant factor in judgments of grade-worthiness and why inconsistent scripts were always rated as the most difficult to grade.

Heuristics may help to explain another comparator that markers have been found to use: recently marked work. For instance, an enquiry which used think aloud

to investigate the marking of experienced Advanced Subsidiary and A-Level Geography examiners found that 'comparing the quality of a candidate's work with their previous responses or with another candidate's work occurred at least once per script on average' (Crisp 2008, 256). Similar findings were reported by Vaughan (1992) who noted that seven of the nine markers in her study judged in this way. Milanovic, Saville and Shuhong (1996, 106) also observed that some markers appeared to judge the level of compositions by comparing 'to the previous one marked'. It would be misleading to describe this process as norm-referencing as the terms of reference are too narrow and too immediate (i.e. the candidate's own previous answers or those of other recently assessed individuals) for norm-referencing which is based on group norms within the wider population. Even the term cohort-referenced may be too wide-ranging to describe the process. The availability heuristic may provide a more meaningful explanation. Availability is a general purpose heuristics and refers to 'cognitive availability', for instance, the ease with which a particular outcome can be pictured (Sherman et al. 2002, 98) or the ease with which instances or associations are called to mind (Tversky and Kahneman, 2002). Laming (2004, 9) also observed that: 'All judgments are comparisons of one thing with another … the judgment depends on what comparator is available'. Because markers typically assess batches of answers rather than single items, they mark under conditions which favour deployment of the availability heuristic.

A role for affect?

A final aspect of marking which may be illuminated by a study of the mind is the role of emotion. The official discourse of professional bodies espouses an image of assessment as an impersonal activity, unclouded by emotion (e.g. Quality Assurance Agency for Higher Education [QAA] 2006). However, markers' verbal reports

frequently attest to the range and strength of emotional response that is triggered by the act of marking. It is easy to see how feelings formed during previous encounters could colour a teacher's assessment of known students. It is, therefore, more revealing to find that emotion remains salient even when marking takes place under experimental conditions where awarding or withholding marks will have no consequences for 'students'. Experimental markers have been found to respond positively when they are able to award marks and negatively when obliged to withhold them. '"Whoopee!"' exclaimed one on finding that an answer was correct; '"Lovely"', remarked another (Suto and Greatorex 2008, 222 and 220). There is palpable relief in another marker's exclamation: '"Phew, that seems to be OK"', on discovering that a candidate had included mark-gaining details (Morgan 1996, 363). These exclamations suggest a desire to award marks, a desire which is also apparent in remarks like: '"I'm hoping for forty-seven"' and '"Would like to give it something but … Pity … it's not in the mark scheme so reluctantly zero"' (Suto and Greatorex 2008, 220 and 225). Morgan's (1996, 361) markers exhibited 'discomfort' when they were obliged to judge harshly, adopting various coping strategies. For instance, they 'shift the blame to an anonymous authority that lays down what "they have to do"'. When the outcome was uncertain, some markers erred on the side of generosity: '"…he hasn't justified it as he's gone along … I probably would give him a 7 though, all the same … OK that's a gut reaction level 7, maybe a bit generous"' (Morgan 1996, 365).

Morgan (1996, 362) argues that the tensions apparent in some of these comments are explained by the conflicting 'positions' adopted by teachers acting as examiners. Her study entailed the use of think aloud and interviews to investigate the marking of GCSE coursework by eleven secondary school mathematics teachers.

Seven different ‘positions’ were identified with markers shifting between them as they struggled to manage these tensions. For instance, an unclear answer forced two teachers to choose between the ‘teacher/advocate’ and the ‘examiner’ position. The ‘teacher/advocate’ position was based on the wish that ‘a pupil should get as high a grade as possible’ and entailed ‘looking for opportunities to give credit’ (361) whilst the ‘examiner’ position entailed the dispassionate application of criteria and a clinical detachment from the author of the work. Although both teachers identified the same features of the answer as significant (the absence of description and explanation), one adopted the teacher/advocate position, arguing that the necessary explanation had probably taken place in the classroom – even though there is no suggestion of this in the text. The other considered the possibility of taking this position but rejected it in favour of the examiner role arguing that the teacher position is only acceptable ‘“if you’re the teacher in the class”’ (361). The different positions led them to ‘opposite rankings’ of the script. Thus, the ‘teacher/advocate’ ranked it highest whilst the ‘examiner’ ‘ranked it lowest of the three texts they read’ (366). Clearly, the positions assumed had considerable consequences for how markers resolved uncertainties about grading.

Other aspects of marking that act as emotional triggers include the content of answers. Thus, one examiner observed: ‘“I am always favourably inclined to a candidate who can interest, surprise, inform or … amuse me”’ whilst another admitted to marking a script down as ‘“trivial – maybe because I know and love the Impressionists”’ (Milanovic, Saville, and Shuhong 1996, 104 and 105). Surface features of performance, including the neatness of presentation and legibility of handwriting, also provoked a response. ‘“Ugh, yuk. What a mess … looking at a whole load of percentages, all over the place”’, exclaimed one marker (Suto and

Greatorex 2008, 225). Poor handwriting was a ‘marked irritation’ to markers in Vaughan’s study (1992, 114) whilst two thirds of Milanovic, Saville, and Shuhong’s (1996, 103) markers ‘seemed to be affected to varying degrees by the handwriting’ with one reporting that ‘“Large, clear writing cheered me up”’.

Markers have also been observed reaching out to the author behind an answer, striving to read the student in their work. This suggests that as well as an emotional aspect to assessment, there is an interpersonal dimension. Again, this has been witnessed in experimental as well as in operational settings. For instance, Morgan (1996, 367) noted how one experimental marker treated a candidate: ‘as an individual with an existence outside the text. Although the text was the only evidence available to her’. Thus, she speculated ‘about what might have happened in the class or what Richard might have done if he had been advised differently’. Similarly, Wyatt-Smith and Castleton (2005, 146) found that teachers marking anonymous work samples persisted in ‘trying to read the student in the writing’, searching, for instance, for traces of gender in the writing.

Recent research in psychology and neuroscience reinforces the testimony of markers, suggesting that these social and affective dimensions play a more fundamental role in judgment than has traditionally been acknowledged, either by research (Hardman 2009, 184) or in the official discourse on assessment as conducted by professional and awarding bodies. Neuroscientists have challenged the received wisdom that judgment is necessarily compromised by emotion by showing how difficult judgment is for individuals who are unable to use prior emotional learning to guide their decision-making. Immordino-Yang and Damasio (2007, 3) argue that: ‘Modern biology reveals humans to be fundamentally emotional and social creatures … It is not that emotions rule our cognition, nor that rational thought does not exist’.

Instead, they argue that emotions 'are profoundly intertwined with thought' (4). Thus, although 'rational thought and logical reasoning do exist', they are 'hardly ever truly devoid of emotion' (7-8). Psychologists in the heuristics and biases tradition have reached similar conclusions. A recently proposed heuristics, the affect heuristic, 'describes the importance of affect in guiding judgments' (Slovic et al. 2002, 397). Its proponents argue for the primacy of affect, claiming that 'affective reactions to stimuli are very often the first reactions, occurring automatically and subsequently guiding information processing and judgment' (398). Whilst this claim to primacy is not universally accepted (e.g. Rottenstreich and Shu 2004), recent research in different disciplines concurs that emotion plays a more complicated and organic role in judgment than has generally been acknowledged.

Conclusion

Marking is a subject of concern across educational sectors. A recent QAA report (2009, 3) noted that its own audit and review reports: 'typically make more recommendations linked to assessment than to any other area'. The programme of work announced at the inauguration of Ofqual illustrates how concerns have extended into the public domain (Tattersall 2008, 9). Thus, Ofqual is undertaking work on the reliability of tests, examinations and teacher assessment, a principal aim of which is to allay public concerns by enhancing understanding of the levels of reliability it is realistic to expect. Matters of concern include the extent to which award standards are being maintained – a concern fuelled by suggestions of 'grade inflation' in GCSE, A-Level and degree awards – and apparent inaccuracies in the application of mark schemes. Most of these wider concerns can be traced directly to the microcosmic level – to judgments made by individuals about specific performances. Thus, it is the contention of this paper that as long as attention is focused on the outward

manifestations of marking, but with insufficient attention to the judgment processes underpinning them, a fundamental component of these difficulties is being overlooked.

The growing corpus of knowledge on marker judgment could make a substantial contribution to the discourse on marking, informing debate and elucidating policy and practice. Whilst the practical implications of the findings discussed above require a separate paper, it is nevertheless important to acknowledge that they do raise questions about established practices. For instance, the application of criteria-based assessment is questioned by a number of the studies cited above. Indeed, many aspects of policy and practice may benefit from review in the light of what is known about judgment. Moreover, this research not only raises questions about existing practice; it may also hold the key to improvements. Various studies have led to advances in practice (e.g. Szpara and Wylie 2005; Suto and Greatorex 2008). Yet the fact remains that only a fraction of the insights that have been yielded by a study of the workings of the mind have been applied to educational assessment. Thus, securing a better understanding of the role of judgment in marking remains the immediate priority and a necessary precursor to the improvement of practice.

REFERENCES

Baddeley, A. 1998. Working memory. *Life Sciences* 321: 167-173.

Bridges, D., J. Elliott, and A. McKee. 1995. *Competence-based Higher Education and the Standards Methodology*. Norwich, University of East Anglia: the Employment Department and the Engineering Council.

Chapman, G. B. 2005. The psychology of medical decision making. In *Blackwell Handbook of judgment and decision making*, eds. D. Koehler and N. Harvey, 585-603. Oxford: Blackwell.

Cooksey, R., P. Freebody, and C. Wyatt-Smith. 2007. Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation* 13: 401-434.

Crisp. V. 2008. Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education* 38: 247-264.

Dennis, I., S. Newstead, and D. Wright. 1996. A new approach to exploring biases in educational assessment. *British Journal of Psychology* 87: 515-534.

Ecclestone, K. 2001. 'I know a 2:1 when I see it': Understanding criteria for degree classifications in franchised university programmes. *Journal of Further and Higher Education* 25: 301-313.

Elander, J. 2004. Student assessment from a psychological perspective. *Psychology Learning and Teaching* 3: 114-121.

Elander, J., and D. Hardman. 2002. An application of judgment analysis to examination marking in psychology. *British Journal of Psychology* 93: 303-328.

Eraut, M. 2000. Non-formal learning and tacit knowledge in professional work. *British Journal of Educational Psychology* 70: 113-136.

Ericsson, K. A. 2002. Towards a procedure for eliciting verbal expression of non-verbal experience without reactivity: Interpreting the verbal overshadowing effect within the theoretical framework for protocol analysis. *Applied Cognitive Psychology* 16: 981-987.

Evans, J. 1993. Bias and rationality. In *Rationality: Psychological and philosophical perspectives*, eds. K. Manktelow, and D. Over, 6-30. London: Routledge.

Fessey, C. 2002. Capturing expertise in the development of practice: Methodology and approaches. *Learning in Health and Social Care* 1: 47-58.

Garry, J., M. McCool, and S. O'Neill. 2005. Are moderators moderate? Testing the 'anchoring and adjustment' hypothesis in the context of marking Politics exams. *Politics* 25: 191-200.

Gilovich, T., and D. Griffin. 2002. Introduction – heuristics and biases: Then and now. In *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin, and D. Kahneman, 1-17. Cambridge: Cambridge University Press.

Gipps, C., and P. Murphy. 1994. *A fair test? Assessment, achievement and equity*. Buckingham: Open University Press.

Grimley, M., H. Dahraei, and R. Riding. 2008. The relationship between anxiety-stability, working memory and cognitive style. *Educational Studies* 34: 213-223.

Hall, K., and A. Harding. 2002. Level descriptions and teacher assessment in England: Towards a community of assessment practice. *Educational Research* 44: 1-15.

Hardman, D. 2009. *Judgment and decision making: Psychological perspectives.* Chichester: BPS Blackwell.

Harlen, W. 2004. *A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes*. In Research Evidence in Education Library. EPPI-Centre, London: Institute of Education.

Harlen, W. 2005. Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education* 20: 245-270.

Hartley, J., M. Trueman, L. Betts, and L. Brodie. 2006. What price presentation? The effects of typographic variables on essay grades. *Assessment and Evaluation in Higher Education* 31: 523-534.

Hay, P., and D. Macdonald. 2008. (Mis)appropriations of criteria and standards-referenced assessment in a performance-based subject. *Assessment in Education: Principles, Policy and Practice* 15: 153-168.

Immordino-Yang, M., and A. Damasio. 2007. We feel, therefore we learn: The relevance of affective and social neuroscience to Education. *Mind, Brain and Education* 1: 3-10.

Jackson, P. 2002. The constructs in people's heads. In *The role of constructs in psychological and educational measurement*, ed. H. Braun, D. Jackson, and D. Wiley. Mahwah, NJ: Lawrence Erlbaum.

Jonsson, A., and G. Svingby. 2007. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review* 2: 130-144.

Keren, G. and K. Teigen. 2004. Yet another look at the heuristics and biases approach. In *Blackwell handbook of judgment and decision making,* ed. D. Koehler, and N. Harvey, 89-109. Oxford: Blackwell.

Laming, D. 2004. *Human judgement: The eye of the beholder*. London: Thomson.

Leighton, J. P. 2004. Avoiding misconception, misuse and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice* 23: 6-15.

Marshall, B. 2000. *English teachers - the unofficial guide: Researching the philosophies of English teachers*. London: RoutledgeFalmer.

Meier, S., B. Rich, and J. Cady. 2006. Teachers' use of rubrics to score non-traditional tasks: Factors related to discrepancies in scoring. *Assessment in Education: Principles, Policy and Practice* 13: 69-95.

Milanovic, M., N. Saville, and S. Shuhong. 1996. A study of the decision-making behaviour of composition markers. In *Studies in Language Testing 3: Performance testing, cognition and assessment,* ed. M. Milanovic, and N. Saville, 92-114. Cambridge: Cambridge University Press.

Morgan, C. 1996. The teacher as examiner: The case of Mathematics coursework. *Assessment in Education: Principles, Policy and Practice* 3: 353-374.

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

Murphy, R.J.L. 1979. Removing the marks from examination scripts before re-marking them: Does it make any difference? *British Journal of Educational Psychology* 49: 73-78.

Newton, P. 1996. The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal* 22: 405-420.

Rottenstreich, Y. and S. Shu. 2004. The connections between affect and decision making: Nine resulting phenomena. In *Blackwell handbook of judgment and decision making,* ed. D. Koehler, and N. Harvey, 444-463. Oxford: Blackwell.

Pollitt, A., and N. Murray. 1996. What raters really pay attention to. In *Studies in Language Testing 3: Performance testing, cognition and assessment,* ed. M. Milanovic, and N. Saville, 74-91. Cambridge: Cambridge University Press.

Price, M., and C. Rust. 1999. The experience of introducing a common criteria assessment grid across an academic department. *Quality in Higher Education* 5: 133-144.

Quality Assurance Agency for Higher Education (QAA). 2006. *Code of practice for the assurance of academic quality and standards in higher education*. Gloucester: Quality Assurance Agency for Higher Education.

Quality Assurance Agency for Higher Education (QAA). 2009. *Thematic enquiries into concerns about academic quality and standards in higher education in England: Final report.* Gloucester: Quality Assurance Agency for Higher Education.

Sadler, D.R. 2005. Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education* 30: 175-194.

Scharaschkin, A., and J. Baird. 2000. The effects of consistency of performance on A Level examiners' judgements of standards. *British Educational Research Journal* 26: 343-358.

Sherman, S., R. Cialdini, D. Schwartzman, and K. Reynolds. 2002. Imagining can heighten or lower the perceived likelihood of contracting a disease: The mediating effect of ease of imagery. In *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin, and D. Kahneman, 98-102. Cambridge: Cambridge University Press.

Slovic, P., M. Finucane, E. Peters, and D. MacGregor. 2002. The affect heuristic. In *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin, and D. Kahneman, 397-420. Cambridge: Cambridge University Press.

Suto, I., and J. Greatorex. 2008. What goes through an examiner's mind? Using verbal protocols to gain insight into the GCSE marking process. *British Educational Research Journal* 34: 213-233.

Szpara, M., and C. Wylie. 2005. National Board for Professional Teaching Standards assessor training: Impact of bias reduction exercises. *Teachers College Record* 107: 803-841.

Tattersall, K. 2008. Speech at the Ofqual launch event. National Motorcycle Museum, Solihull, UK, 16 May.

Tversky, A., and D. Kahneman. 2002. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. In *Heuristics and biases: The psychology of intuitive judgment*, ed. T. Gilovich, D. Griffin, and D. Kahneman, 19-48. Cambridge: Cambridge University Press.

Van Someren, M., Y. Barnard, and J. Sandberg. 1994. *The think aloud method: A practical guide to modelling cognitive processes*. London: Academic Press.

Vaughan, C. 1992. Holisitic assessment: What goes on in the rater's mind? In *Assessing second language writing in academic contexts*, ed. L. Hamp-Lyons, 111-125. Norwood, NJ: Ablex.

Wenger, E. 1998. *Communities of practice: Learning, meaning and identity*. Cambridge: Cambridge University Press.

Wiliam, D. 1992. Some technical issues in assessment: A user's guide. *British Journal of Curriculum and Assessment* 2: 11-20.

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

Wilmut, J., R. Wood, and R. Murphy. 1996. *A review of research into the reliability of examinations: Discussion paper prepared for the School Curriculum and Assessment Authority*. Nottingham: University of Nottingham.
http://www.nottingham.ac.uk/education/centres/cdell/pdf-reportsrelexam/relexam.pdf (accessed February 5, 2003).

Wolf, A. 1995. *Competence-based assessment*. Buckingham: Open University Press.

Wyatt-Smith, C., and G. Castleton. 2005. Examining how teachers judge student writing: An Australian case study. *Journal of Curriculum Studies* 37: 131-154.