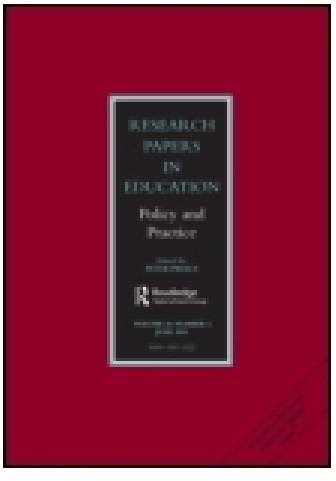

Research Papers in Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rred20

On the reliability of high-stakes teacher assessment

Sandra Johnson [a]
[a] Assessment Europe and University of Bristol , Bristol , UK
Published online: 18 Jan 2013.

To cite this article: Sandra Johnson (2013) On the reliability of high-stakes teacher assessment, Research Papers in Education, 28:1, 91-105, DOI: 10.1080/02671522.2012.754229

To link to this article: http://dx.doi.org/10.1080/02671522.2012.754229

Research Papers in Education, 2013
Vol. 28, No. 1, 91–105, http://dx.doi.org/10.1080/02671522.2012.754229

On the reliability of high-stakes teacher assessment

Sandra Johnson*

Assessment Europe and University of Bristol, Bristol, UK

For a number of reasons, increasing reliance is being placed on teacher assessment in high-stakes contexts in many countries around the world. Simultaneously, countries that have for some time relied to greater or lesser degrees on teacher assessment for high-stakes purposes are in the process of questioning the validity of that reliance. In principle, teacher assessment has an important role to play in increasing assessment validity by complementing testing to cover subject domains more comprehensively than otherwise would be possible. But what is the evidence regarding the reliability of teacher assessment in high-stakes contexts? The answer is that the evidence is limited and often ambiguous. Research has revealed that teachers can be influenced by a number of construct-irrelevant factors as they work towards their judgements, factors such as gender, socio-economic background, effort and behaviour, that risk biasing their assessments. And when considering construct-relevant achievement evidence teachers are often expected to use verbal or semi-verbal sets of criteria, such as level descriptions, which typically require a degree of subjective interpretation in application and so are themselves a source of unwanted variation in judging standards. Arguably the most effective strategy for addressing these issues is participation in consensus moderation. Yet there have been few attempts to provide evidence of the effectiveness of moderation in practice. The potential value of, and the growing reliance upon, teacher assessment in high-stakes applications demand that evaluation of consensus moderation become a built-in part of the process.

Keywords: summative; high-stakes; teacher assessment; moderation; reliability

Introduction

The literature on teacher assessment is vast, often controversial and sometimes contradictory. This paper does not set out to review the entire question of teacher assessment but, in keeping with the theme of this volume, restricts attention to the issue of the *reliability* of teacher assessment, and, in particular, to evidence in support of claims for and against the reliability of teacher assessment in high-stakes contexts (other forms of teacher assessment, in particular formative assessment in the classroom, are not considered here). For the most part, discussion focuses on public examinations in the UK, although reference is also made to other countries where practice and experience appear relevant to the UK situation.

*Email: Sandra.Johnson@Assessment-Europe.com

There are a number of reasons for appealing to teacher assessment in place of or alongside test-based assessment in public examinations and other high-stakes assessment systems. Principal among them are:

- to provide assessments of abilities and skills that are not amenable to testing, which if not available would reduce the validity of the overall subject assessment;
- to exploit the rich base of evidence that teachers have available to them for making assessments by virtue of the time spent interacting with and observing their students, that could in principle lead to greater validity and reliability even when tests might be used to assess the same abilities and skills;
- to minimise any potentially disruptive impact that formal testing might have on coursework and learning, along with any psychologically damaging effects that tests and frequent testing might have on students;
- to respect the professionalism of teachers, to empower them and to help them maintain a high level of assessment skill in both summative and formative assessment;
- and, under certain conditions, to minimise the cost of large-scale assessment.

Teacher assessment has a particularly important role to play in complementing test-based assessment, enabling assessment to more fully span the intended subject domain than it otherwise might. To quote Stanley et al. (2009):

> … the teacher is increasingly being seen as the primary assessor in the most important aspects of assessment. The broadening of assessment is based on a view that there are aspects of learning that are important but cannot be adequately assessed by formal external tests. These aspects require human judgment to integrate the many elements of performance behaviours that are required in dealing with authentic assessment tasks. (Stanley et al. 2009, 31)

When formal tests could be used with complete validity to assess a knowledge domain or a particular set of subject skills, then one could reasonably ask under what circumstances teacher assessment might be a better alternative. Where they are applicable, tests can, in principle, have some advantages over teacher assessment. In particular, the nature and range of the knowledge and skills being assessed are usually clear from a review of test content and mark scheme, as is the relative importance being given to various different aspects of learning or to different intellectual or personal qualities. The degree of construct validity offered by a test is in consequence relatively transparent. Reliability can also, in principle, be empirically investigated – though it might be noted that reliability is a relatively neglected area in testing as in other forms of assessment, a phenomenon that has only recently begun to be addressed in England by the Office of Qualifications and Examinations Regulation (Ofqual) – see Opposs and He (2012).

While teacher assessment does not lend itself to the same degree of transparency, it might nevertheless be the preferred option. Teachers are with their students for extensive periods of time, constantly interacting with them inside and sometimes outside the classroom, posing questions and noting responses, and observing performances as they carry out assigned tasks and activities. In consequence, teachers are assumed to have a more comprehensive view of their students as learners and

achievers than any set of test results can alone provide, arguably leading not only to potentially higher validity than time-constrained tests can embody (Wiliam 2001, 2003; Harlen 2007), but also, some believe, to potentially higher reliability than tests can achieve (Wiliam 2001, 2003; Daugherty 2011). Moreover, the fewer the number of formal tests that students are required to attempt the less disruption there will be to normal classroom schedules and the less anxiety students will experience (Harlen and Deakin Crick 2002, 2003; Harlen 2004a).

An important motivation for using teacher assessment in place of, or in addition to, tests is to recognise, develop and value the professionalism of teachers. One negative impact of the higher profile given to test-based results in England's national curriculum assessment system has been shown to be not only a loss of assessment skill on the part of teachers, but also a loss of confidence in their ability to make sound assessments of their students (Black et al. 2010, 2011). Indeed, so little value has in practice been given to teacher judgement in this system at key stage 2 in England (end of primary schooling) that the window for submitting level judgements has overlapped the date by which test results are delivered to the schools; test results alone have been used as indicators of pupil, school and system performance. Addressing this issue, it was recommended in a recent official review that teacher assessments should in future be submitted *before* test results are known by schools (Bew 2011), giving teacher assessment greater credibility. It remains to be seen whether this intention will be achieved.

Finally, there is the issue of cost-effectiveness. Any large-scale high-stakes assessment system that can function as effectively on the basis of teacher assessment as it could on the basis of a formal testing infrastructure could be a more cost-effective system, particularly where teachers offer their assessments freely. How important any saving might be, or if there might be a saving at all, will, however, depend on the financial investment put into quality assurance through assessment literacy development for teachers, school support, monitoring and moderation: the state of Queensland in Australia provides teachers and schools with extensive, and expensive, support in efforts to quality assure the judgements of teachers in the statewide senior secondary certification system (QSA 2010).

There can be little doubt that teachers represent a wealth of knowledge about students' achievements and capabilities that is indispensable in the assessment of learning progress and achievement, and which, *in principle*, could usefully be exploited in high-stakes examination and certification systems. For this latter purpose, however, it is essential that the assessments that teachers make of their students can not only be shown to be valid but are also demonstrated to be sufficiently reliable for purpose. So what is the evidence regarding the reliability of teacher assessment in high-stakes contexts?

Teacher assessment reliability: the sparse evidence

The legitimacy of teacher assessment in high-stakes assessment systems depends critically on the degree to which 'teacher exchangeability' can be assumed. Should a piece of achievement evidence be rated by one teacher rather than another, what is the likelihood that the rating outcome will be the same? And would this likelihood change should the teachers concerned be located in different schools, different authorities, different regions and different states?

Where teachers mark a traditional test using a given mark scheme then the effects on candidate outcomes of marker-related factors can be formally evaluated using conventional reliability estimation approaches (see, for example, Meadows and Billington 2005 for a review of marker reliability studies in the UK, and Newton and Meadows 2011). Similar studies could, in principle, be carried out for situations where tangible achievement evidence other than completed tests is the subject of teacher rating, using grade-related descriptors; such studies would include the evaluation of art creations, oral performances, science investigations and so on. In practice, multiple-rater studies in these teacher assessment contexts are rare. In their place are various hybrids, in particular comparing teachers' marks with those of an acknowledged 'expert' marker (the human benchmark, who is assumed to carry 'the truth' in mark value terms) or their qualitative judgements with the single or consensus judgements of one or two expert raters ('moderators', 'verifiers', 'panellists'), or comparing teachers' marks or judgements with the marks or classifications produced by a test.

To summarise, among the different strategies that have been employed in attempts to explore the reliability of teacher assessment are the following:

(1) comparing the independent assessments of different teachers (inter-rater reliability studies);
(2) comparing the assessments of teachers with those of 'expert' raters (human benchmark exercises);
(3) comparing teachers' assessments with test results (teacher-test agreement studies).

Given the importance of the issue, several reviews have confirmed that astonishingly few reports of investigations into the reliability of teacher assessment are to be found in the literature, particularly as regards high-stakes applications (Harlen 2004b; Wilmut 2005; Stanley et al. 2009; Johnson 2012), a situation almost paralleling that for tests. Here are some examples.

Teacher assessment of coursework has been a contributing feature in the General Certificate of Secondary Education (GCSE) examining system in England, Wales and Northern Ireland since the GCSE itself was launched in the late 1980s, and plays a role also in many subject examinations in the General Certificate of Education (GCE) Advanced Subsidiary (AS) and Advanced (A) levels.[1] Over time, a number of concerns began to emerge about the quality of this component in subject examinations, particularly in the GCSE. Of particular relevance to this paper was the fear that both validity and reliability were threatened by teachers' use of inappropriate assessment tasks, by their misapplication of assessment criteria, by the application of differing marking standards, and by pupil cheating and plagiarism. In response to a review by the regulator, the Qualifications and Curriculum Authority (QCA 2006), the uncontrolled coursework assessment in GCSE examinations was replaced with 'controlled assessment'. Teacher assessment now complements test-based assessment rather than duplicating or substituting for it, and where it continues to contribute to a subject examination, its weighting is 25% or 60%. Awarding bodies exercise varying degrees of control over the tasks teachers use in their assessment, over the conditions in which the tasks are used and the assessments made, and over the evaluation of the resulting assessment evidence (see QCDA[2] 2009; Johnson 2012, chapter 3).

In both the GCSE and GCE examination systems, the quality of coursework assessment across schools is, in principle, checked through a system of judgemental moderation, in which awarding organisation moderators – practising subject teachers – review small samples of students' work from each school, and either agree that the marks awarded are appropriate or that some adjustment, even an entire re-mark, is warranted (see Johnson 2012, chapter 4, for an overview). A single moderator essentially verifies the standards of any particular school. Disappointingly, and surprisingly, given the importance of the teacher-assessed elements in some subject examinations, there have been no multiple-rater studies, or at least no published multiple-rater studies, into the reliability of coursework assessment since the early 1990s. In a rare study (Taylor 1992), three experienced teacher moderators independently evaluated the coursework portfolios of 60–80 candidates for each of four internally assessed examination components. In GCSE English, 15–25% of candidates would have been awarded a different grade had their teacher's mark been replaced by the mark of one or other of the moderators; in GCSE history the proportion varied between approximately 20% and 40%; in GCSE mathematics the variation was 15–30%; in A level psychology it was 10–20%. This was an important piece of work that has not, to the author's knowledge, been repeated in the intervening 20 years.

With no relevant reliability evidence available from designed inter-rater studies, can we glean anything useful from a comparison of teachers' grade predictions with grades actually achieved in GCSE and GCE subject examinations? It is common practice for teachers to submit subject grade predictions for their students to the relevant awarding organisation ahead of the examinations. The resulting grade distributions are used each year as one element in the battery of information that informs standard setting procedures that lead to grade awarding decisions (Robinson 2007; Johnson 2012, chapter 4). Studies have consistently shown teacher-examination grade agreement rates to be moderate to low, with some variation in the strength of relationship related to the subject of the qualification (Dhillon 2005). High agreement rates could have been considered as validity and reliability evidence for both grading systems; moderate to low agreement rates offer no support one way or the other for either.

In the absence of any relevant evidence about the reliability of teacher assessment in the high-stakes examination systems in the UK, let us look further afield.

Every state in Australia incorporates an element of internal assessment in its senior secondary certification system (Cumming and Maxwell 2004). The state of Queensland has for over 30 years been operating a school leaving certificating system that is *entirely* dependent on teacher assessment. The Queensland system reports five broad levels of achievement ('very high achievement' down to 'very limited achievement') in a number of different subjects, each level containing 10 'exit rungs'. Quality assurance is given high priority in a multi-tier system of checks. The process begins with accreditation of school plans for implementing subject syllabuses, followed by different stages of monitoring, verifying and approving standards of judgement, and ends with a check on judgement consistency statewide through a post-certification review of random samples of assessed folios (for details see Maxwell 2006; QSA 2010). In the random sampling checks, district panel members work in pairs, independently assessing assigned folios before arriving at a consensus classification judgement. These panel judgements are then compared with the schools' original decisions. The latest random sampling exercise confirmed previous

years' findings, reporting that 87% of the reviewed folios were placed by the reviewers at the same achievement level as the schools, with some variation across subjects; 'serious disagreement' was recorded in almost 4% of cases (QSA 2011, 1). Where outcomes differed, panellists tended more often to award lower classifications to students than their schools did (the district panel review takes place after the schools' certification decisions have been announced, so that the review results can only feed into the following year's rating practices).

New South Wales offers another Australian example. Here, test scores for basic and advanced examination papers in English and mathematics are complemented by statistically moderated teacher assessments in the state's well-established high-stakes assessment system. Interestingly, after decades of operation, there has only recently been growing awareness that little is actually known, and certainly little published, about the underlying reliability of the teacher assessments before they are statistically adjusted. One response has been an analysis of a set of data furnished over five years (2004–2008) of statewide assessment (MacCann and Stanley 2010). In each subject, the distributions of raw test scores and of moderated teacher marks were first converted to percentile grade distributions. Then, the degree of match in grade classifications was computed separately for the paired test results, the paired teacher assessments and the paired composites. The results were interpreted as showing teacher assessment to be slightly more reliable than test results, with composite score classifications slightly more reliable than either set alone. However, the validity of these inferences depends on the assumption that the two teacher assessments, like the two test results, were *independent* measures. This assumption can readily be challenged, since teachers rating their own students for basic and advanced achievement are likely to be influenced to some degree by the well-known 'halo effect', putting into question the validity of the reported finding about the superiority of teacher judgements over test results.

In the absence of other reports on teacher assessment reliability in situations that are high stakes for students, we look now at evidence from other contexts.

In national curriculum assessment at key stage 2 in England, an assessment system that is high stakes for teachers and schools if not for students, both teachers and tests report subject attainment in the form of level classifications within a range of four progression levels (for a comprehensive overview, see Johnson 2011, chapter 5). A recent study compared teachers' level judgements with level classifications made on the basis of a reading pretest (Hutchison and Benton 2010, chapter 11). The exact agreement rate was found to be 66%, a figure lower than that reported for an earlier study that reported agreement rates of around 75% for live test classifications against teacher judgements for reading, mathematics and science (Reeves, Boyle, and Christie 2001). An important difference between the two studies is that in the later study, teacher and test classifications were independent, whereas in the earlier study, an unknown proportion of teachers could have been offering their judgements in full knowledge of the live test results (the teacher judgement submission window spanning delivery of test results to schools). In both studies, it was impossible to infer from the results whether teacher judgements were more reliable than test classifications or vice versa.

An interesting example of an over-time reliability study, also involving teacher assessment at key stage 2, is reported in Stanley et al. (2009, 44). In the study, teachers' level judgements for over 10,000 pupils over seven consecutive years, provided both during pretests and at the time of live testing just weeks later, were

analysed. A principal finding was that in just over 80% of cases the teachers' judgements for individual pupils exactly agreed on the two occasions. However, as is the case for all studies in which any measureable time has elapsed between assessments, what this result actually tells us about the consistency with which teachers can rate their students is unclear. Does it indicate a high degree of consistency in teachers' judgements, with only one in five teachers coming to a different view on the two occasions? Or could it be that the consistency rate would have been even higher if some students had not changed their state of 'levelness' over the period? Indeed, how many of the students given the same rating on both occasions actually merited that rating both times?

A study in Scotland produced even lower agreement rates against a level-based progression framework that most teachers had been using in that country for over a decade (Hayward 2007). Teachers' level judgements were compared with test results for large samples of eight to 14-year-old pupils, using data furnished by the Scottish Survey of Achievement (SSA) for reading, mathematics and science (Johnson and Munro 2008). Rates of exact level agreement were highest for numeracy/mathematics, at between 45% and 60% per age group, followed by reading, with around 40% agreement for all groups, and with science showing the lowest rates of agreement at between 10% and 35%. The researchers offered several possible explanations for the particularly poor agreement rates in science, including a likely disparity in the nature of the science being assessed by teachers and tests, viz. 'process' vs. 'knowledge'. In other words, teachers and tests might in science have been assessing quite different constructs. Unfortunately, no research has been carried out before or since that might clarify the issue.

A number of studies have explored the reliability of teacher assessment of students' writing skills. National curriculum assessment in England again offers an example. With single-marking of scripts the norm, with no evidence available about the reliability of this marking, and with markers difficult to find in the UK to redress any potential problem, the QCA (the then regulator) commissioned a specially designed reliability study, in which groups of English and Australian teachers marked key stage 3 reading and writing scripts (Baker et al. 2008). Important group differences were noted when marks were converted to 'levels', and compared with the pre-assigned level classifications of an 'expert marker'. The English markers' results agreed with the expert for 57% of the reading scripts and 56% of the writing scripts, with corresponding figures of 68% and 41%, respectively, for the Australian markers. The import of these findings for inferences about rater reliability is unclear, given the reliance on comparisons of markers with the single expert marker (the human benchmark) rather than with each other, as in a genuine inter-rater reliability study.

An SSA-related quality assurance study carried out in Scotland (Johnson 2010) was specifically designed to estimate the reliability of writing assessment, using the replication-based definition of reliability. Put simply, interest here was in the difference that might be expected should one classroom teacher rather than another rate a particular student's piece of writing. A number of primary teachers independently rated pieces of extended writing produced by 8–14-year olds, arriving at level judgements after applying a scheme of national writing criteria that had been in use in schools for over a decade. Analysis revealed that rater–script interaction was a larger contributor to measurement error than differences in rater severity. In other words, while markers, as might be expected, differed to some degree in their overall

standards of rating, differences were not systematic but varied across scripts. In the associated national assessment programme, in which three teachers independently judged each piece of submitted writing, there was majority agreement on level in around 90% of cases and unanimous agreement in around one-third of cases (see, e.g. SSA 2006); reliability coefficients for the 2009 survey were over 0.85 (Johnson 2010).

In the writing reliability studies described above, the rating was of one single piece of writing from each student. The only possible source of measurement error that could be explored, therefore, was the rating of that piece by the teacher raters. But it is increasingly recognised that just as students typically vary in their performance across the items in a traditional subject test so, too, can their writing performance vary depending on the writing task they are set. An example of a study in which several pieces of writing per student were independently evaluated by several different raters is described by Schoonen (2005). Of particular interest here is the fact that uneven performances across the students' multiple writing tasks, i.e. student–task interaction, contributed more to assessment unreliability than did differences between raters or rater–task interaction. This same finding has emerged in other areas where studies have been able to explore reliability using several tasks and several raters, including the assessment of students' science investigation skills (Shavelson, Baxter, and Gao 1993) and medical students' performance in clinical diagnosis (Murphy et al. 2009). Such performance variability across items or tasks is an important interaction effect that contributes to measurement error.

The potential – albeit unproven – role of moderation in assuring reliability

Within high-stakes systems, teacher assessment takes many forms, from systematically marking test papers and assignments using provided mark schemes to arriving at 'global' assessments of subject achievement at the end of an entire academic year or longer period of work. The subject matter might be clearly or quite loosely defined, and observable evidence of achievement might be permanent, as in a completed test paper, a piece of writing, an art work, or a portfolio, 'virtually' permanent, as in a video recording of an oral presentation or interchange, or ephemeral, as in performance in a drama production or sports event (see Johnson 2012 for an overview of the situation in English school leaving examinations).

It is unlikely that the same degree of assessment reliability will be achievable for all of these different contexts and types of achievement evidence – indeed, the extent to which reliability is even an accessible concept in some cases can be questioned. But even in the absence of knowledge about the reliability achieved we can identify a number of factors that will threaten it (Wilmut 2005, section 7; Johnson 2012, chapter 5, summarised in Baird et al. 2012, 816–818). Contributions to unreliability include inadequate support structures, varying construct perceptions on the part of the teacher assessors, lack of clarity in and applicability of assessment criteria, and reliance on insufficient or inappropriate bases of achievement evidence for evaluation.

Despite guidance, different teachers might operationalise their assessments of a subject domain differently, perhaps giving greater emphasis to some aspects than others: for example, investigation skills vs. factual knowledge in science, computational skills vs. graphical skills in numeracy, or grammar vs. style in writing. The evidence teachers use to arrive at their assessments might also vary in important

ways, in terms of its nature and extent. Even when construct perception is shared and the evidence base is similar, teachers might judge the evidence in different ways, using different criteria and applying different standards. Little research has been carried out to investigate this issue, as noted by Harlen (2004a) in her review of teacher assessment reliability studies, despite the undoubted impact on reliability of clarity in criteria; the dearth of relevant research has only recently begun to be addressed (Wyatt-Smith and Castleton 2005; Black et al. 2010, 2011; Wyatt-Smith, Klenowski, and Gunn 2010).

In addition, teachers have been shown to be consciously or unconsciously influenced by construct-irrelevant student characteristics (Harlen 2004a, 2005; Wyatt-Smith and Castleton 2005; Martinez, Stecher, and Borko 2009; Wyatt-Smith, Klenowski, and Gunn 2010). Among these are gender (Lafontaine and Monseur 2009; Ready and Wright 2011), ethnicity (Burgess and Greaves 2009; Ready and Wright 2011), socio-economic status (Hauser-Cram, Sirin, and Stipek 2003; Wyatt-Smith and Castleton 2005; Ready and Wright 2011), special educational needs status (Thomas et al. 1998; Reeves, Boyle, and Christie 2001) and personality traits, in particular behaviour and effort (Bennett et al. 1993; Morgan and Watson 2002; Wyatt-Smith and Castleton 2005; Wyatt-Smith, Klenowski, and Gunn 2010). Unfortunately, unless such biases are systematic and student numbers are large, it is impossible to quantify the effect of these types of influence on the quality of assessment.

Statistical moderation of teacher assessments, as practised in the high-stakes assessment system in operation in New South Wales, can address issues of gross differences in overall standards of judgement between schools, districts and regions, using test data as a reference point. But there are two issues to note here. The first is that if a test is needed for the purpose of adjusting teacher assessments, then why have teachers make the assessments in the first place? If the aim of involving teachers in this extra workload is in some sense to value their professionalism, increasing their confidence and self-esteem along with developing their ability to assess their students dependably both for formative and summative purposes, then how is that aim achieved? The second issue is that adjusting school-based teacher assessment distributions to match those of tests might adequately address between-school, and even between-class, differences in assessment standards, but it cannot address the equally important question of teacher–student interaction effects.

Post-certification benchmarking of the sort used in Queensland has two weaknesses. The first is that, since the review takes place after the event, any injustices seen to be done to individual students by their teachers' assessments remain unidentified, or, if identified, unaddressed. The second is that while panellist members are, in principle, acting as benchmarks against which teachers' assessments are being judged, there is no evidence that panellist pairs are themselves interchangeable as carriers of standards (the same comment applies to expert markers in test-based contexts). If they cannot be proven to be interchangeable then the validity of their review judgements is open to question.

Potentially, the most effective strategy for ensuring both validity and reliability in teacher assessment, if these can in principle be achieved to an acceptable degree, is consensus moderation (Hutchinson and Hayward 2005; Wyatt-Smith, Klenowski, and Gunn 2010; Black et al. 2011). This is the process by which teacher assessors are brought together to consider construct definition and assessment criteria, and jointly to review, discuss and judge student work samples, with a view to acquiring

a shared understanding of standards; the process is common practice in large New South Wales secondary schools, but not systematically practised in the UK. Merely implementing such a moderation system, though, is insufficient for purpose. Proof of efficacy should be required, in the form of empirical evidence of sustainable inter-rater reliability, not only locally but also regionally and nationally. Reliability needs not only to be demonstrated during or immediately following consensus moderation events but later too, to establish that out-of-context experience can successfully be transferred to day-to-day practice. On what other basis should governments and others be expected to agree to provide and support the necessary infrastructural resources (Daugherty 2007, 2011) that would be needed to roll out such a moderation system nationwide for a high-stakes assessment purpose?

Regrettably, despite the need for such proof, studies that have offered any relevant evidence are extremely rare. Even recent small-scale studies that have specifically set out to explore the ways that teachers make their assessments within a consensus moderation context have not added that final inter-rater reliability study (Wyatt-Smith, Klenowski, and Gunn 2010; Black et al. 2011). One research study that did evaluate the impact of moderation on assessment reliability is England's Monitoring Children's Progress Project, the general objective of which was to facilitate confidence and competence in teacher assessment within the national curriculum assessment programme. Towards the end of the three-year experiment, empirical data began to emerge showing that teachers could reach a reasonably high rate of agreement in their level judgements (Stanley et al. 2009 offer full evaluation details). A second study is the Scottish writing assessment reliability study referred to earlier (Johnson 2010). One objective of this study was to evaluate the impact on teachers' level judgement agreement rates of their participation in a typical one-day consensus moderation exercise involving around 30 teachers. The meeting began with an overview both of writing as a construct and of the assessment criteria. This was followed by a collective review with discussion and rating of exemplar scripts. The finding was that while their experience in the moderation meeting was much valued by the teachers (a frequently reported outcome in the literature), their assessment behaviour was no different after participation than it had been before. Did the moderation format need a redesign or was it that the teachers concerned were already so familiar with the rating scheme and its use that they had no need for formal consensus moderation?

Discussion

As Stanley and colleagues have remarked in the context of teacher assessment reliability:

> there is considerable interest in getting an indication of what levels of reliability and validity are commonly obtained in similar situations. (Stanley et al. 2009, 38)

Unfortunately, we are a very long way from knowing whether any generalities hold for the levels of reliability that might be achieved in any given type of situation. This is partly because the variety of situations is very large, and the threats to reliability are diverse. It is also a consequence of the dearth of relevant research evidence at this time.

Apart from the sparsity of reported empirical studies in this area, the other striking finding from reviews of teacher assessment reliability (Harlen 2004b; Wilmut

2005; Stanley et al. 2009; Johnson 2012) must be the fact that so few of the small number of published studies have emanated from organisations responsible for providing and reporting the results of high-stakes certification systems in which teachers' summative assessments play some part. This is equally true for vocational assessment.

There is a widely felt desire within the educational community at large to value teachers' professional expertise, and, in the UK, an associated belief in the potential of teacher summative assessment to right the perceived wrongs of a controversial and unpopular testing regime that permeates the length of schooling. An appeal to moderation, it is assumed, will ensure validity, consistency and comparability in teachers' assessments. Sadly, there is at present no convincing evidence to support this assumption. Indeed, the 2011 review of the key stage 2 testing system 'heard evidence from a number of assessment experts about the limits of moderation for making teacher assessment judgements more robust and reliable' (Bew 2011, 50).

In Wales, where the national curriculum testing regime was abandoned in 2005 in favour of teacher assessment for all purposes, including accountability, the inspectorate has recently expressed concern that to date no empirical evidence has been made available about the effectiveness of moderation in assuring assessment reliability nationwide (Estyn 2010). In Australia, where moderated teacher assessment has been given a high-profile role over many years in student certification, there has apparently been little empirical research into issues of underlying validity and reliability: 'Teacher judgement in Australia remains largely uncharted territory and legitimate influences on judgement need to be investigated' (Wyatt-Smith, Klenowski, and Gunn 2010, 61). In Sweden, which abandoned school-leaving examinations in favour of teacher assessment in the 1990s, there is growing concern about the wisdom of that move, in terms of the likely dependability of teacher assessment results (Wikstrom 2006; Gustafsson and Erickson 2011); one issue, as in Scotland, is evidence of grade inflation over time in an atmosphere of school competitiveness (Wikstrom 2006). In Scotland and the Netherlands, where national qualifications involve mixtures of test-based and internal assessment, no published evidence is yet available about the reliability of internal assessment, principally because it is difficult, even impossible, to provide it (van Rijn, Béguin, and Verstralen 2012).

In any high-stakes context, a dependence on teacher assessment of unknown quality is as undesirable as dependence on tests of questionable validity and reliability, even where the teacher assessment addresses aspects of achievement that tests cannot. Before adopting any system of assessment for high-stakes purposes, it is essential that the dependability of the intended assessment methodology be investigated and confirmed as fit for purpose. A mere belief in the superiority of one approach over another is not sufficient. The degree of validity and reliability achieved should be established before system implementation, and not left to chance, as it has been in recent years in all countries of the UK, and elsewhere, with quite predictable consequences.

It is essential to provide teachers with clear assessment criteria and guidelines, and consensus moderation is a necessity, at least in the early stages of implementation of a new rating scheme. However, consensus moderation can only prove its worth if formally evaluated. Both formative and summative evaluation should be designed into the process to include empirical evidence of the degree to which standards of judgement can as a result be shared among teachers. Evaluation would

include the production of inter-rater reliability indices using generalizability theory (Brennan 2001; Cardinet, Johnson, and Pini 2010) or other approaches (in particular classification agreement rates), in time for design modifications to be implemented and the process re-evaluated as often as necessary to achieve and to maintain the level of effectiveness demanded by assessment system stakeholders. There would clearly be significant financial and logistic costs associated with the inclusion of a formal evaluation element in moderation systems, but the gains in knowledge about assessment dependability would surely justify these.

Acknowledgements

This paper owes much to the constructive feedback on earlier drafts of the two special issue guest editors and the three anonymous reviewers.

Notes

1. The GCSE examinations are taken at the statutory school-leaving age of 16. The GCE comprises the AS and A-level examinations, which are taken during the next two years of schooling. The AS is a lower level, taken on average in four subjects, whilst A is a higher level, taken on average in three subjects which will already have been amongst those studied at AS level.
2. QCDA was the Qualifications and Curriculum Development Authority, successor to the QCA.

Notes on contributor

Sandra Johnson is a visiting fellow in the Graduate School of Education, University of Bristol, and director of Assessment Europe, a company offering technical support in educational assessment. Her research interests include assessment quality, teacher assessment, grade comparability and qualification equivalence. Recent publications include: *Assessing learning in the primary classroom* (2011); *A focus on teacher assessment reliability in GCSE and GCE* (2012). Qualifications and mobility in a globalising world: Why equivalence matters. *Assessment in Education* (2009).

References

Baird, J.-A., A. Beguin, P. Black, A. Pollitt, and G. Stanley. 2012. "The Reliability Programme Final Report of the Technical Advisory Group." In *Ofqual's Reliability Compendium*, edited by D. Opposs and Q. He, 771–838. Coventry: Office of Examinations and Qualifications Regulation.

Baker, E. L., P. Ayres, H. F. O'Neil, K. Choi, W. Sawyer, R. M. Sylvester, and B. Carroll. 2008. *KS3 English Test Marker Study in Australia. Final report to the National Assessment Agency of England*. London: National Assessment Agency.

Bennett, R. E., R. L. Gottesman, D. A. Rock, and F. Cerullo. 1993. "Influence of Behaviour Perceptions and Gender on Teachers' Judgements of Students' Academic Skill." *Journal of Educational Psychology* 85: 347–356.

Bew, P. 2011. *Independent Review of Key Stage 2 testing, assessment and accountability.* Final Report to the British Government Department for Education.

Black, P., C. Harrison, J. Hodgen, B. Marshall, and N. Serret. 2010. "Validity in Teachers' Summative Assessments." *Assessment in Education* 17: 215–232.

Black, P., C. Harrison, J. Hodgen, B. Marshall, and N. Serret. 2011. "Can Teachers' Summative Assessments Produce Dependable Results and Enhance Classroom Learning?" *Assessment in Education* 18: 451–469.

Brennan, R. L. 2001. *Generalizability Theory.* New York, NY: Springer-Verlag.

Burgess, S., and E. Greaves. 2009. *Test Scores, Subjective Assessment and Stereotyping of Ethnic Minorities*. Working paper 09/221, University of Bristol, Centre for Market and Public Organization.

Cardinet, J., S. Johnson, and G.-R. Pini. 2010. *Applying Generalizability Theory using EduG*. New York, NY: Routledge.
Cumming, J. J., and G. S. Maxwell. 2004. "Assessment in Australian Schools: Current Practice and Trends." *Assessment in Education* 11: 94–108.
Daugherty, R. 2007. "National Curriculum Assessment in Wales: Evidence-informed Policy?" *Welsh Journal of Education* 14: 62–77.
Daugherty, R. 2011. "Designing and Implementing a Teacher-based Assessment System: Where is the infrastructure?" Paper presented at the Oxford University Centre for Educational Assessment seminar *Teachers' judgments within systems of summative assessment: strategies for enhancing consistency*, Oxford, June.
Dhillon, D. 2005. "Teachers' Estimates of Candidates' Grades. Curriculum 2000 Advanced Level Qualifications." *British Educational Research Journal* 31: 69–88.
Estyn. 2010. *Evaluation of the Arrangements to Assure the Consistency of Teacher Assessment in the Core Subjects at Key Stage 2 and Key Stage 3*. Cardiff: Her Majesty's Inspectorate for Education and Training in Wales.
Gustafsson, J.-E., and G. Erickson. 2011. "To Trust or Not to Trust? Contrasting Findings from Teachers' Assessments." Paper presented at the annual conference of the Association for Educational Assessment – Europe, Belfast, Northern Ireland, November.
Harlen, W. 2004a. *A Systematic Review of the Evidence of the Impact on Students, Teachers and the Curriculum of the Process of Using Assessment by Teachers for Summative Purposes*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
Harlen, W. 2004b. *A Systematic Review of the Evidence of the Reliability and Validity of Assessment by Teachers for Summative Purposes*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
Harlen, W. 2005. "Trusting Teachers' Judgements: Research Evidence of the Reliability and Validity of Teachers' Assessment Used for Summative Purposes." *Research Papers in Education* 20: 245–270.
Harlen, W. 2007. *Assessment of Learning*. London: Sage.
Harlen, W., and R. Deakin Crick. 2002. *A Systematic Review of the Impact of Summative Assessment and Tests on Students' Motivation for Learning*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.
Harlen, W., and R. Deakin Crick. 2003. "Testing and Motivation to Learn." *Assessment in Education* 10: 170–207.
Hauser-Cram, P., S. R. Sirin, and D. J. Stipek. 2003. "When Teachers' and Parents' Values Differ: Teacher Ratings of Academic Competence in Children from Low-income Families." *Journal of Educational Psychology* 95: 813–820.
Hayward, E. L. 2007. "Curriculum, Pedagogies and Assessment in Scotland: The Quest for Social Justice. 'Ah kent yir faither'." *Assessment in Education* 14: 251–268.
Hutchinson, C., and L. Hayward. 2005. "The Journey so Far: Assessment for Learning in Scotland." *The Curriculum Journal* 16: 225–248.
Hutchison, D., and T. Benton. 2010. *Parallel Universes and Parallel Measures: Estimating the Reliability of Test Results*. Coventry: Office of Qualifications and Examinations Regulation.
Johnson, S. 2010. *The Reliability of Writing in the 2009 Survey*. Internal report produced for the Scottish Government.
Johnson, S. 2011. *Assessing Learning in the Primary Classroom*. London: Routledge.
Johnson, S. 2012. "A Focus on Teacher Assessment Reliability in GCSE and GCE." In *Ofqual's Reliability Compendium*, edited by D. Opposs and Q. He, 365–416. Coventry: Office of Qualifications and Examinations Regulation.
Johnson, S., and L. Munro. 2008. "Teacher Judgements and Test Results: Should Teachers and Tests Agree?" Paper presented at the Annual Conference of the *Association of Educational Assessment – Europe*, Hissar, Bulgaria, November.
Lafontaine, D., and C. Monseur. 2009. "Les évaluations des performances en mathématiques sont-elles influencées par le sexe de l'élève?" *Mesure et Evaluation en Education* 32: 71–98.

MacCann, R. G., and G. Stanley. 2010. "Classification Consistency When Scores are Converted to Grades: Examination Marks Versus Moderated School Assessments." *Assessment in Education* 17: 255–272.
Martinez, J. F., B. Stecher, and H. Borko. 2009. "Classroom Assessment Practices, Teacher Judgments, and Student Achievement in Mathematics: Evidence from the ECLS." *Educational Assessment* 14: 78–102.
Maxwell, G. 2006. "Quality Management of School-based Assessments: Moderation of Teacher Judgements." Paper presented at the 32nd IAEA Conference, Singapore, May.
Meadows, M., and L. Billington. 2005. *A Review of the Literature on Marking Reliability*. London: National Assessment Agency.
Morgan, C., and A. Watson. 2002. "The Interpretive Nature of Teachers' Assessment of Students' Mathematics: Issues for Equity." *Journal of Research in Mathematics Education* 33: 78–110.
Murphy, D. J., D. A. Bruce, S. W. Mercer, and K. W. Eva. 2009. "The Reliability of Workplace-based Assessment in Postgraduate Medical Education and Training: A National Evaluation in General Practice in the United Kingdom." *Advances in Health Sciences Education* 14: 219–232.
Newton, P., and M. Meadows. 2011. "Special Issue: Marking Quality Within Test and Examination Systems." *Assessment in Education*, 18: 213–216.
Opposs, D., and Q. He, eds. 2012. *Ofqual's Reliability Compendium*. Coventry: Office of Qualifications and Examinations Regulation.
QCA. 2006. *A Review of GCSE Coursework*. London: Qualifications and Curriculum Authority.
QCDA. 2009. *Changes to GCSEs and the Introduction of Controlled Assessment for GCSEs*. London: Qualifications and Curriculum Development Agency.
QSA. 2010. *Moderation Handbook for Authority Subjects*. Brisbane: Queensland Studies Authority.
QSA. 2011. *Random Sampling Project. 2011 Report on Random Sampling of Assessment in Authority subjects*. Brisbane: Queensland Studies Authority.
Ready, D. D., and D. L. Wright. 2011. "Accuracy and Inaccuracy in Teachers' Perceptions of Young Children's Cognitive Abilities: The Role of Child Background and Classroom Context." *American Educational Research Journal* 48: 335–360.
Reeves, D. J., W. F. Boyle, and T. Christie. 2001. "The Relationship Between Teacher Assessments and Pupil Attainments in Standard Test Tasks at Key Stage 2, 1996–98." *British Educational Research Journal* 27: 141–160.
Robinson, C. 2007. "Awarding Examination Grades: Current Processes." In *Techniques for Monitoring the Comparability of Examination Standards*, edited by P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, and P. Tymms, 97–123. London: Qualifications and Curriculum Authority.
Schoonen, R. 2005. "Generalizability of Writing Scores: An Application of Structural Equation Modelling." *Language Testing* 22: 1–30.
Shavelson, R. J., G. P. Baxter, and X. Gao. 1993. "Sampling Variability of Performance Assessments." *Journal of Educational Measurement* 30: 215–232.
SSA. 2006. *Scottish Survey of Achievement. 2005 English language and Core Skills* – Practitioner's Report. Edinburgh: Scottish Government.
Stanley, G., R. MacCann, J. Gardner, L. Reynolds, and I. Wild. 2009. *Review of Teacher Assessment: Evidence of What Works Best and Issues for Development*. Oxford: University of Oxford Centre for Educational Assessment.
Taylor, M. 1992. *The Reliability of Judgements Made by Coursework Assessors*. Associated Examining Board internal report.
Thomas, S., G. E. Madaus, A. E. Raczek, and R. Smees. 1998. "Comparing Teacher Assessment and Standard Task Results in England: The Relationship Between Pupil Characteristics and Attainment." *Assessment in Education* 5: 213–246.
van Rijn, P. W., A. A. Béguin, and H. H. F. M. Verstralen. 2012. "Educational Measurement Issues and Implications of High Stakes Decisions Making in Final Examinations in Secondary Education in the Netherlands." *Assessment in Education* 19: 117–136.
Wikstrom, C. 2006. "Education and Assessment in Sweden." *Assessment in Education* 13: 113–128.

Wiliam, D. 2001. "Validity, Reliability and all that Jazz." *Education* 3–13 (29): 17–21.
Wiliam, D. 2003. "National Curriculum Assessment: How to Make it Better." *Research Papers in Education* 18: 129–136.
Wilmut, J. 2005. *Experiences of Summative Teacher Assessment in the UK*. London: Qualifications and Curriculum Authority.
Wyatt-Smith, C., and G. Castleton. 2005. "Examining How Teachers Judge Student Writing: An Australian Case Study." *Journal of Curriculum Studies* 37: 131–154.
Wyatt-Smith, C., V. Klenowski, and S. Gunn. 2010. "The Centrality of Teachers' Judgement Practice in Assessment: A Study of Standards in Moderation." *Assessment in Education* 17: 59–75.