A comparative judgement approach to teacher assessment

Suzanne McMahon

Ashbourne School, County Meath

Ian Jones

Mathematics Education Centre, Loughborough University

Author note

Suzanne McMahon, Ashbourne School, County Meath, Ireland; Ian Jones, Mathematics Education Centre, Loughborough University, UK.

Correspondence concerning this article should be addressed to Ian Jones, Mathematics Education Centre, Ann Packer Building, Loughborough University, Loughborough, LE11 3TU, UK. Email: I.Jones@lboro.ac.uk

Abstract

We report one teacher's response to a top-down shift from external examinations to internal teacher assessment for summative purposes in the Republic of Ireland. The teacher adopted a comparative judgement approach to the assessment of secondary students' understanding of a chemistry experiment. The aims of the research were to investigate whether comparative judgement can produce assessment outcomes that are valid and reliable without producing undue workload for the teachers involved. Comparative judgement outcomes correlated as expected with both test marks and with existing student achievement data, supporting the validity of the approach. Further analysis suggested that teacher judgement privileged scientific understanding, whereas marking privileged factual recall. The estimated reliability of the outcome was acceptably high, but comparative judgement was notably more time consuming than marking. We consider how validity and efficiency might be improved, and the contributions that comparative judgement might offer to summative assessment, moderation of teacher assessment, and peer assessment.

Keywords: Teacher assessment, summative assessment, comparative judgement, chemistry

Introduction

The research reported here was motivated by one teacher's response to top-down changes to the Junior Cycle component of the national assessment system in the Republic of Ireland (DES, 2012).

Changes to the Junior Cycle include newly developed subjects and short courses, a focus on literacy, numeracy and key skills, and broad approaches to assessment and reporting. A key intention is that schools will have more freedom to design programmes that meet the learning needs of all students (NCCA, 2009), and the most significant change is in the area of assessment. Final assessments set by the State Examinations Commission will continue to play a role, contributing to 60% of a given student's final outcome, but administered and marked by schools. The remaining 40% of a student's outcome will be allocated to school-based assessments carried out by subject-specialist teachers during the second and third years of the Junior Cycle.

Analogous changes are evident around the world, such as the current removal of national standardised "levels" in favour of greater teacher assessment in England, Wales and Northern Ireland (DfE, 2013). Conversely, in jurisdictions where teacher assessment has traditionally been a major part of summative assessment, questions are being raised about its reliability (e.g. Jones, 2013), and systems of external standardised assessment are being introduced (e.g Hume, 2014).

In this paper we report an innovative approach to implementing a school-based assessment that sought to maximise the potential benefits of teacher assessment while minimising potential drawbacks. We first review the literature on teacher assessment within the context of changes to the Irish assessment system and the reported study. We then describe the comparative judgement approach to assessment that was adopted for the study, and set out our motivation for applying it to summative teacher assessment. Following this we report the main study and discuss the findings in terms of validity, reliability and implications for teacher workload.

Teacher assessment

The research literature offers a broad consensus on the potential promises and drawbacks of using teacher assessment for summative purposes (ARG, 2006; Johnson, 2013). In the remainder of this section we summarise the consensus, focusing on issues of validity, reliability and teacher workload.

Validity. Validity is widely debated in the literature (e.g. Lissitz, 2009; Messick, 1994; Wiliam, 2001), and is taken here to refer to the fitness of the assessment for categorising students in terms of their learning achievement within a given domain. Our focus is on the extent to which the test used in the study generated student responses that evidenced achievement in chemistry, and the extent to which the judging procedure adopted utilised the evidence to produce valid outcomes.

A key motivation for using teacher assessment for summative purposes is to improve the validity of what is assessed (Johnson, 2013). The current Junior Certificate exams have been criticised for encouraging an "emphasis on rote learning and on rehearsing questions" (DES, 2012, p.1). Teacher assessment might allow the broadening of summative assessment to include "a wide variety of methods, tasks and strategies" (p.18), including authentic and performance-based assessments that are difficult or impossible to capture using traditional exams (DES, 2012; QCA, 2009). However, the research evidence on the validity of teacher assessment in practice is mixed. While some reviews have found that teacher assessment outcomes correlate strongly with achievement data and standardised assessments overall (Hoge & Coladarci, 1989), others have found this is not always the case (Harlen, 2004; Stanley, MacCann, Gardner, Reynolds & Wild, 2009).

One reported threat to validity is teacher bias based on broader knowledge of students (Bennett, Gottesman & Rock, 1993; Hoge & Butcher, 1984; Wilson & Wright, 1993). In the study reported here we aimed to minimise bias through anonymisation, and by distributing the assessment across teachers who were not familiar with all the students. Another reported threat to validity is teachers' lack of training and expertise in designing and conducting summative assessment activities (Black, Harrison, Hodgen, Marshall & Serret, 2010; Choi, 1999). We attempted to minimise this problem using a comparative judgement approach, as described in the following section, which has been shown to be successful with untrained assessors (Jones & Alcock, 2013; Jones & Wheadon, submitted).

Reliability. Reliability is the extent to which outcomes would be replicated if an assessment were carried out again (Harlen, 2004). For example, a reliable test would result in similar responses if it were re-administered to the same students. A reliable assessment procedure would result in the same outcomes if student responses were re-assessed by an independent group of examiners. Other comparative judgement studies have reported high inter-rater reliabilities ranging from .79 to .95 (Jones & Alcock, 2013; Jones, Inglis, Gilmore & Hodgen, 2013; Jones, Swan & Pollitt, 2014). In the present study, practical constraints prevented re-administering the test or reassessing it using an independent group of teachers. However, estimation of the scale separation of the outcomes provided us with indirect evidence regarding the inter-rater reliability of the procedure.

Improved reliability of national assessments has been cited as an incentive to use teacher judgement for summative purposes (MacCann & Stanley, 2010; Wiliam, 2001).[1] Wiliam (2001) argued that ongoing teacher assessments could be analogous to increasing the length of standardised tests, thereby increasing reliability. However, it is not clear that in practice this would necessarily be achieved due to the paucity and inconclusiveness of evidence about the reliability of teacher assessment (Johnson, 2013; Stanley et al., 2009; Tisi, Whitehouse, Maughan & Burdett, 2013).

One potential threat to reliability arises from poor design of assessment tasks by teachers (Harlen, 2004). In this study we describe the performance of the assessment used, which included a relatively unstructured item designed to evidence students' scientific understanding of a chemistry practical. Another threat to reliability is the varying standards of judgement across teachers (Cross & Frary, 1999; Wyatt-Smith, Klenowski & Gunn, 2010). Johnson (2013) recently argued for consensus moderation in high-stakes teacher assessment to address this variability. Here we adopted a comparative judgement approach that pools individual assessment decisions to produce a collective assessment outcome (Pollitt, 2012). In this sense, the approach adopted here is inherently moderated from the outset.

Teacher workload. A third motivation for using teacher assessment for summative purposes is that it might free teachers and students from the burden of preparing for external examinations (ARG, 2006; Crooks, 1988; Johnson, 2013). Moreover, it

[1] We note that the term "reliability" does not appear in official documentation about changes to summative assessment in the Junior Cycle (DES, 2012; NCCA, 2009).

might provide more opportunity for assessment to be integrated with teaching and learning (Black & Wiliam, 2009; Wiliam, 2001; Harlan, 2004). However, the incoming changes to the Junior Cycle in Ireland are highly contentious, with a particular concern that support structures are not in place and so teacher workload will dramatically increase (Donnelly, 2014). In this paper we do not address the politics or controversy surrounding the current policy changes. Rather, we report one teacher's attempt to validly and reliably assess scientific understanding without introducing undue workload.

In the following section we detail the comparative judgement approach adopted for the teacher assessment exercise, and describe how it might help minimise the above threats to validity, reliability and teacher workload.

Comparative judgement

Comparative judgement originates with psychophysical research in the 19th Century, and was developed in the 20th Century, most notably by Thurstone (1927). It has more recently been developed for educational assessment purposes by Pollitt (2012). It is based on holistic judgements of pairs of students' test responses by a group of assessors who make independent judgments as individuals. For each pairing of test responses an assessor must decide which has the greater of a specified global construct, in our case "scientific understanding". The outcome is a binary decision matrix of the "winner" and "loser" for each pairing. The matrix is then fitted to a statistical model, such as the Bradley-Terry model (Bradley & Terry, 1952; Firth, 2005), to produce a parameter estimate and standard error of the "value" of each test response. The parameter estimates can be used to construct a scaled rank order of test responses from "best" to "worst", and used for the usual assessment purposes such as grading.

Comparative judgement approaches have been applied to a variety of educational purposes over recent years. These include examination comparability studies (Bramley, 2007; Jones et al., 2014), peer assessment of undergraduate mathematics (Jones & Alcock, 2013), teacher assessment of creative writing (Heldsinger & Humphry, 2010), design and technology e-portfolios (Kimbell, 2012; Newhouse, 2014), and practical science (Davies, Collier & Howe, 2012). These studies have

consistently yielded reliable and valid assessment outcomes across a variety of contexts where estimates have been reported.

The psychological rationale for comparative judgement is that human beings are competent when comparing one object against another, but unreliable when rating objects in isolation (Gill & Bramley, 2013; Laming, 1984; Thurstone, 1927). When a teacher marks student work using a rubric they draw on their knowledge of other students' work (Crisp, 2013). Marking practices attempt to minimise this comparative influence through detailed and specific rubrics (Pollitt, 2004). Comparative judgement takes the opposite approach, harnessing the comparative aspect of assessment directly and dispensing with rubrics and marking.

Comparative judgement offers design options that might be preferable to marking procedures in certain contexts. We argue that teacher assessment for summative purposes can provide one such context. In the remainder of this section we set out how comparative judgement might minimise the threats to validity, reliability and workload summarised above.

Bias. One source of bias arises from teachers' knowledge and expectations of students influencing their judgements. Comparative judgement might help reduce bias by pooling classes and teachers for summative assessment purposes. In the study reported here five teachers judged test responses from five classes, and so were unlikely to be familiar with all the students assessed.

The tests were also anonymised by removing student names and replacing them with a unique indentification number prior to assessment activities. Anonymisation is not particular to comparative judgement, but is impractical for many forms of teacher assessment. However, it is possible that teachers inferred the identity of some of their own students from handwriting and presentation.

Other sources of bias, such as favouring neater handwriting or other construct-irrelevant features (Husbands, 1976; Johnson, 2013) are likely to have been present. While such biases are common to many assessment approaches, we acknowledge that they might be particularly problematic for the types of open-response questions suited to comparative judgement.

Training. Many scholars advocate professional development in assessment theory and practice prior to introducing teacher assessment (Choi, 1999). The study reported here

was in one sense a professional development exercise in itself for the teachers involved, but one initiated without prior training. This was a practical decision due to time constraints. Nevertheless, comparative judgement offers a method for quickly implementing innovative assessments. In previous studies we have found psychometrically sound results without developing detailed items or marking rubrics, and using untrained assessors (Jones et al., 2013; Jones & Alcock 2013; Jones & Wheadon, submitted).

Assessment tasks. A key design challenge for teachers using comparative judgement is selecting prompts or tasks that can generate varied student responses relevant to the construct of interest. Variation in the student responses generated is required such that teachers are able to discern a difference when judging pairs of test responses. Moreover, teachers will only be able to make valid judgements if all the responses provide evidence relevant to the construct. We return to this design challenge in the analysis and discussion sections.

Teacher standards. Variation in teacher judgements of student achievement can be reduced in two ways. First, people are more consistent when making relative judgements rather than absolute judgements (Laming, 1984; Thurstone, 1927). We would expect teachers to be more consistent when asked if test response A is better than test response B, than if asked to give each test response a mark out of ten. The comparative judgement approach can therefore be expected to reduce inconsistencies in teacher assessment outcomes.

Second, comparative judgement should be undertaken by a group of assessors and not a lone assessor. While we assume a broad homogeneity across assessors' interpretation of a specified global construct, we can assume some variation too. For example, one teacher may privilege formal scientific vocabulary, and another may privilege accurate diagrams. Moreover, some judges may have a better grasp of the construct, some may be less distracted by surface features, and so on. The comparative judgement approach is grounded in assessors' collective decisions, and so individuals' preferences are balanced and their biases are diluted. This is not the case with traditional marking where each test response is normally marked by just one teacher.

Consistency between teachers is ideally explored by estimating the inter-rater reliability. Two groups of teachers drawn from the same population independently assess the test responses and the outcomes are compared. However, this was not directly possible in the present study due to time and other constraints. Instead an internal reliability measure, the Scale Separation Reliability (SSR) as described later, was used to provide an estimate of the level of "agreement" between teachers when judging the test responses (Pollitt, 2012). The SSR is analogous to estimates of inter-rater reliability, for example as might be obtained from double-marking in the traditional approach.

Teacher workload. A traditional barrier to using comparative judgement for educational purposes has been the number of judgement decisions required to construct accurate parameter estimates for each test response. The study reported here involved 154 test responses and the total possible number of unique pairings is 11,781.[2] However, research has shown that satisfactory scale separation can be produced using just a small sample of possible pairings (Bradley & Terry, 1952; Pollitt, 2012). In this study the five participating teachers made 1550 pairwise decisions between them, just 13.2% of the total possible number of unique pairings.

Nevertheless the efficiency of comparative judgement remains a key challenge to its implementation. A key question is whether each teachers' allocation of 310 pairwise judgements takes more time than marking about 31 test responses each. Clearly, for the time to be equivalent each teacher's allocation must be 10 pairings in the time it would take to mark a single test response. Although this seems onerous comparative judgement assumes fast, intuitive decisions (Pollitt, 2012). Moreover each test response was seen on average about twenty times across the five teachers. As such a key contrast between marking and comparative judgement is whether each test response is scrutinised in detail once, or judged quickly many times. This has implications for the appropriateness of assessment tasks for the judging as compared to marking, and we return to this issue in the discussion.

A separate issue regarding teacher workload is the extent to which summative assessment innovations might complement learning and teaching, rather than adding work unrelated to classroom practice. Previous work suggests that comparative

[2] The number of unique pairings for n test responses is calculated using $\frac{1}{2} \times n(n-1)$.

judgement lends itself to peer assessment activities in which students judge one another's work (Jones & Alcock, 2013; Jones & Wheadon, submitted). We investigated this possibility as an adjunct to the main study.

Research aims

The broad research aim of the study was to explore the theoretical and practical implications of a comparative judgement approach to summative teacher assessment in secondary school chemistry.

The theoretical focus was mainly on the validity of the assessment outcomes in terms of existing achievement data, independent marking and the coherence and scale separation of the test used. We also indirectly investigated the inter-rater reliability of the outcomes by considering the scale separation reliability produced by the teachers' judgments.

The practical focus was mainly on the efficiency of the method, and the use of the outcomes for grading purposes. We also explored the potential for integrating comparative judgement with teaching and learning practices by conducting a peer assessment exercise.

Method

Participants. One hundred and fifty-four students aged 14 and 15 years old and five teachers from a large secondary school in the Republic of Ireland participated in the study.

Materials. A teacher (the first author) designed a test, which was a template for writing up an experimental report, as shown in Appendix 1. A marking rubric was also designed, based on an Assessment Marking scheme devised by the State Examinations Commission (http://tinyurl.com/ks533lq), also shown in Appendix 1. Comparative judgement was implemented online using the No More Marking website (www.nomoremarking.com[3]).

[3] The second author is the senior scientific advisor to No More Marking Ltd. which owns the www.nomoremarking.com website.

Procedure. The students completed a scientific investigation on the solubility of potassium chloride in one lesson. In the following lesson they were required to complete the test under examination conditions, and no prior warning was given.

The test responses were anonymised, scanned and uploaded to the No More Marking website. The responses were comparatively judged online by the five teachers over a one-month period. The teachers completed 1550 pairwise judgements which were fitted to the Bradley-Terry model using the statistical software R (Firth, 2005). The outcome was a parameter estimate and standard error[4] for each test response. The parameter estimates were then used to construct a scaled rank order of test responses from "best" to "worst". For more detail about the comparative judgement modelling process see Pollitt (2012) and Bramley (2007).

The test responses were also marked using the rubric, shown in Appendix 1. Two teachers who were involved in the judging marked all 154 responses between them, and the marks for the four sections were aggregated to produce a total mark for each student. The marks ranged from 2 to 31 (out of a possible 31), with a mean 15.71 and a standard deviation 5.83. Marking was undertaken purely as a research activity and both comparative judgement and marking would not ordinarily be undertaken for routine educational use. However, the teacher initiating the innovation felt that demonstrating agreement between marks and parameter estimates would provide a strong and accessible reassurance to colleagues that comparative judgement can yield valid outcomes. Ideally, the marking would have been undertaken independently by teachers not involved in the comparative judgement component of the study, and the test responses would have been marked twice independently to calculate an inter-rater reliability estimate. However, practical constraints forbade these possibilities. Reports of high-stakes examinations in science and mathematics suggest inter-rater reliabilities of up to .99 are achievable (Murphy, 1982; Newton, 1996). For low-stakes research activities involving questions designed for comparative judgement this might be lower. For example, Jones & Inglis (submitted) reported an inter-rater reliability of .91 for the marking of a problem-solving mathematics exam.

To further explore validity, external achievement marks were generated by calculating the mean mark from four chemistry tests (maximum possible mark 100) administered

[4] The standard errors were calculated with 'quasi variances' (Firth, 2003) using the package qvcalc in R.

over the previous two years as part of Christmas and summer exams. One hundred and twenty-eight students' marks were available for all four tests, 21 students' marks were available for between one and three tests, and the remaining five students were excluded from the prior achievement analysis due to no previous marks being available. The external achievement marks ranged from 12 to 92, with mean 56.82 and standard deviation 18.52.

Following this a departmental meeting took place in which the teachers reviewed the assessment outcomes and applied grade boundaries to the scaled rank order. Outliers were identified and the outlier test responses scrutinised.

Finally, 37 students from two of the six classes that participated in the study peer assessed the test responses using a similar comparative judgement procedure. The validity and scaled separation analysis was repeated for the outcomes of the peer assessment exercise.

Analysis and results

There were two main parts to the analysis. First, scale separation was calculated for the judging procedure to obtain an indirect estimate of inter-rater reliability, and the consistency across different judges and test responses was also estimated. Second, validity was investigated by calculating Pearson product-moment correlation coefficients between all the assessment data (teacher judging, teacher marking and prior achievement data). Validity was further explored by conducting multiple linear regression, using individual item marks as predictors for comparative judgement parameter estimates. The analysis was then repeated for the peer assessment outcomes, and the outcome of the students' assessments compared to those of the teachers.

Scale Separation Reliability (SSR) of the comparative judgement outcomes was calculated to provide an indirect estimate of inter-rater reliability (Andrich, 1982). SSR provides a measure of the separatedness of the parameter estimates over the size of the standard errors generated when the decision data is statistically modelled. It is calculated using the following formula

$$SSR = \frac{SD^2 - RMSE^2}{SD^2}$$

where SD is the standard deviation of the parameter estimates and RMSE is the root mean square error of the standard error associated with each test response's parameter estimate. Therefore separation can be increased by greater discrimination between the response (increasing the standard deviation of the estimates), or increasing the number of judgements (and therefore reducing the error associated with each estimate). We found that the scale separation reliability was acceptably high, SSR = .874.

The consistency of judges was also investigated by calculating "misfit" figures, standardised with a mean of 0 and standard deviation of 1, which provide an indication of how consistent each judge was with the overall consensus. The usual guideline is to consider any judge whose misfit figure is greater than two standard deviations above the mean misfit figure to be performing at odds with the other judges (Pollitt, 2012). All the judges were found to be within two standard deviations of the mean misfit figure suggesting the judges were consistent. However, one judge's misfit figure was close to two standard deviations (1.74) and notably different to the figures of the other judges (range -0.72 to -0.11), and may have been judging a slightly different construct, or judging somewhat erratically. Similarly, misfit figures were calculated for the test responses to provide an indication of how consistently each test response was judged across the judges. Eight of the 154 responses' misfit figures were outside two standard deviations of the mean misfit figure, suggesting 94.8% of responses were judged consistently.

To investigate validity, the comparative judgement parameter estimates were correlated with test response marks. The parameter estimates were found to correlate highly with marks, $r = .715$, as shown in Figure 1. The parameter estimates correlated moderately with prior achievement data, which was available for 149 of the 154 students, $r = .536$. This correlation coefficient compares favourably with the mean of the correlations between the marks of the four tests that made up the prior achievement data, $r = .351$. The correlational analyses support the validity of using comparative judgement to assess the test responses, as summarised in Table 1.

TABLE 1 HERE

Insights helpful for interpreting validity and reliability arose at a meeting in which the five teachers reviewed the outcomes and assigned grades to students (see below). During this meeting the teachers identified by eye the four outliers highlighted in Figure 1 (with these test responses removed the correlation between test marks and teacher parameter estimates rises from $r = .715$ to $r = .788$). Three of the outliers were marked moderately to highly but appeared low in the scaled rank order of parameter estimates. Scrutiny of the test responses suggested that these students had obtained most of their marks from questions 1, 2 and 4 of the test (see Appendix 1), which required short objective answers. Conversely question 3 was completed poorly, with one student not writing anything and another appearing to have written up the wrong experiment. The fourth outlier was marked moderately but came out second top in the scaled rank order generated from the teachers' comparative judgments. This student scored poorly on questions 1 and 2 but teachers judged that her method demonstrated a sound scientific understanding.

FIGURE 1 HERE

The analysis of outliers suggested that when making judgement decisions the teachers privileged the quality of responses to question 3, which was relatively open, over the accuracy or completeness of responses to questions 1, 2 and 4, which required short, objective answers. To investigate this further we conducted a multiple linear regression on the parameter estimates using marks for questions 1 to 4 as predictors. The four questions explained 60% of the variance in the parameter estimates, $R^2 = .60$, $F(4, 149) = 56.57$, $p < .001$. Questions 1 to 3 were significant predictors, $\beta_1 = .16$, $p = .013$, $\beta_2 = .19$, $p = .002$ and $\beta_3 = .60$, $p < .001$, but question 4 was not a significant predictor, $\beta_4 = .06$, $p = .271$. As expected, question 3 was the strongest predictor of parameter estimate, and alone explained 52% of the variance, $R^2 = .52$, $F(1, 152) = 166.00$, $p < .001$. This analysis supports the appropriateness of comparative judgement for relatively open items that might better assess scientific understanding, rather than closed items designed to assess factual recall.

Efficiency

Precise timing data for the judging and marking procedures were not available. For each judgement, the time from the initial presentation of a pair of test responses and the judge's decision was recorded. However, this provides at best an over-estimate of judging time because it is likely teachers were at times interrupted or distracted during some judgements. Scrutiny revealed 29 of the 1550 judgement timing estimates were unexpectedly long (greater than five minutes), suggesting that teachers were sometimes interrupted while making an online judgement decision. These timing estimates were removed and the average time per judgement and total time for all judgements was calculated. We found that the average time taken by teachers to make a single pairwise judgement was 33 seconds, and that the total judging time for all judgements was about fourteen hours, just under three hours per teacher. These numbers were consistent with the teachers' own estimates of how long the judging work had taken.

The time taken to anonymise, scan and upload the test responses for online judging should also be considered. Anonymisation was implemented using a removable cover sheet and individualised tests containing a unique identification number. The tests were scanned and uploaded in a single batch. Altogether the process was not very time consuming, taking less than an hour. In addition, a teacher spent a further hour writing the level descriptors shown in Appendix 2.

The two teachers who marked the tests reported that this had taken about one and a half hours each, totaling approximately three hours to mark all 154 test responses. The teachers were instructed only to score the responses and not to add written comments or provide other forms of feedback. Had comments and feedback been incorporated into the marking procedure this would undoubtedly have increased marking time.

Nevertheless, comparative judgement took considerably longer to complete than the marking exercise, challenging the efficiency of the method compared to traditional procedures. It should be noted that the comparative judgement procedure was inherently moderated due to each test response being seen by several different teachers. If the marking had been independently re-marked, this would have doubled the time required to six hours. Nonetheless, the comparative judging took fourteen hours, more than twice as long as even double-marking.

We return to issues of efficiency and how it might be improved in the discussion.

Peer assessment

A peer assessment exercise in which some students comparatively judged the responses was undertaken following the teacher assessments. This was to explore the potential of integrating comparative judgement approaches with teaching and learning.

Thirty-seven students completed between 24 and 200 judgements each (mode = 100) during a science lesson. The assessment was set up such that no student saw her or his own test response. Students worked alone at a computer but were encouraged to discuss the responses and their judging decisions whilst working. The total 3722 judgement decisions were statistically modelled and analytic procedures undertaken similar to those described above.

The scale separation reliability of the peer assessment outcome was acceptably high, SSR = .893, indirectly suggesting good inter-rater reliability. Standardised judge misfit figures were calculated (mean 0, standard deviation 1) as described above. Thirty-five of the 37 judges were within two standard deviations of the mean misfit figure (range -1.40 to 1.36) , suggesting overall consistency in the judging of 94.6% of the students. Of the two misfitting students, one was marginally two standard deviations (2.19), and the other almost four standard deviations (3.95), above the mean. For the test responses, ten misfit figures were outside two standard deviations of the mean, suggesting 93.5% of responses were judged consistently.

To explore validity, the student parameter estimates were correlated with the teacher parameter estimates. The correlation was high, $r = .741$, suggesting good agreement between students and teachers. We also correlated the student parameter estimates with teacher marks for the tests, and again found this to be high, $r = .667$. Moreover the correlation between the student estimates and marks was not significantly different from the correlation between the teacher estimates and marks, Steiger's $Z = 1.21$, $p = .227$. Finally, we correlated the student estimates with prior achievement data and found this to be moderate, $r = .551$, and not significantly different from the correlation between teacher estimates and achievement data, Steiger's $Z = .33$, $p = .739$.

Taken together, these findings suggest the peer assessment exercise produced valid assessment outcomes, as summarised in Table 1. However, the agreement between teachers and students was high but somewhat short of total agreement ($r = .741$). It is therefore possible that students were viewing and attending to the test responses differently to the teachers. To explore this we conducted multiple linear regression on the students' parameter estimates using individual marks for questions 1 to 4 as predictors. For the teacher assessment we found that question 3 was a substantially stronger predictor than the other questions as reported above, and we were interested to see if this was the case also for the peer assessment. The four questions explained 46% of the variance in student parameter estimates, $R^2 = .46$, $F(4, 149) = 31.53$, $p < .001$. As was the case for the teachers, questions 1 to 3 were significant predictors of parameter estimates, $\beta_1 = .29$, $p < .001$, $\beta_2 = .28$, $p < .001$ and $\beta_3 = .28$, $p < .001$, but question 4 was not a significant predictor, $\beta_4 = .10$, $p = .110$. However, questions 1 to 3 were equally strong predictors suggesting that, in contrast to the teachers, the students did not privilege responses to question 3 over the other questions when making their judgement decisions. This suggests that unlike the teachers, the students attended to factual recall (questions 1 and 2) as much as scientific explanation (question 3) when making their judgement decisions.

Grading

The teachers applied grade boundaries to the scaled rank order of test responses, and grades were fed back to the students. Although not of direct theoretical relevance to the focus of the paper, the grading and feedback exercise is of practical interest, and we report it here.

Grading began with scrutiny of the scaled rank order, as shown in Figure 2. The teachers felt that three grade boundaries were sufficient. They began at the high end of the scale, identifying a break point shown just to the left of the middle-high boundary line in Figure 2. Scrutiny of the test responses surrounding the proposed boundary resulted in the boundary being moved to the right by one test response. A similar break was identified for the lower boundary, two test responses to the right of the low-middle line shown in Figure 2. Again scrutiny of the test responses

surrounding the proposed boundary resulted in it being shifted, this time two test responses to the left.

FIGURE 2 HERE

Following this, one of the teachers wrote level descriptors based on the test responses within each boundary (low-middle-high), as shown in Appendix 2. Students received one of three grades and the corresponding level descriptor.

We also considered how differences in the rank orders produced by teacher judgement, marking and student judgement might have impacted upon grade classification. Direct comparison was not possible because grade boundaries had not been applied to the rank orders produced by marking and student judgement. As described above, grade boundaries had not been established normatively and so could not simply be applied across all three rank orders to make a comparison. Nevertheless, to provide an estimate of classification consistency, we applied the grade boundaries decided by the teachers as though they were norm-referenced to the marking and student judgement rank orders. Figure 3 shows the rank order of test responses for the three assessment procedures with each response shaded according to its grade based on teacher judgement. The overall agreement of grades between the rank orders produced by marking and teacher judgement was 66.9%, and the agreement between the rank orders produced by student judgement and teacher judgement was 63.0%.

FIGURE 3 HERE

Discussion

The comparative judgement approach to teacher assessment as reported here was found to produce valid outcomes. Parameter estimates correlated as expected with traditional marks and with prior achievement data. The resulting scale had good separation reliability, indirectly suggesting that the inter-rater reliability was acceptably high. The correlation of teacher assessment outcomes with those of

students provides further support to the validity and reliability of the assessment outcomes.

The comparative judgement procedure was less efficient than traditional marking in terms of the time required by the teachers. Related to the issue of efficiency, we demonstrated how, in principle, comparative judgement might be integrated with learning and teaching activities through students judging the test responses.

In the remainder of the discussion we reflect on the findings in terms of validity and efficiency, and how these measures might be maximised in future implementations. We also consider other ways in which comparative judgement might be applied to teacher assessment, namely moderation purposes and to enhance student learning.

Validity. The criterion validity of the comparative judgement procedure was supported by correlations with marking and existing achievement data, as well as the estimated grading classification consistency, as summarised in Table 1. Moreover, the multiple linear regression analysis, using individual question marks to predict comparative judgement parameter estimates, suggested that teachers based their judgements largely on performance on question 3, which was relatively unstructured. The other three questions, which assessed recall of key words, may therefore have introduced noise into the judging outcomes.

The multiple-regression outcome is consistent with a study into the use of comparative judgement to measure students' conceptual understanding of mathematics. Jones et al. (2013) administered a two-part test that required students to rearrange a list of fractions in order of size, and then explain their reasoning. Analysis revealed that general mathematics achievement grades were significant predictors of comparative judgement parameter estimates, but marks for the ordering of fractions part of the test were not. The authors concluded that the judges privileged conceptual (explanatory) over procedural (computational) aspects of test responses when making pairwise decisions.

An analogous judging process may have operated in the present study. The teachers appeared to privilege explanatory understanding of the experiment over factual recall. Therefore, omitting questions 1, 2 and 4, and extending question 3, might have improved the outcomes. More generally, tests designed for comparative judging might produce more valid and reliable outcomes when they consist of an open item

that requires students to explain their understanding in a sustained manner, and exclude closed items that require rote recall or the application of procedures.

Furthermore, the test used here was relatively traditional. A promising aspect of comparative judgement is for the summative assessment of work that does not lend itself to reliable marking. Such work includes essays (Heldsinger & Humphry, 2010), ePortfolios of practical work (Kimbell, 2012; Newhouse, 2014), and conceptual understanding of mathematics (Jones & Alcock, 2013). Such tests also tend not to make use of items that require factual recall, thereby avoiding the problem highlighted by the multiple-regression analysis. Further work is required to establish the reliability, validity and efficiency of such approaches.

Efficiency. A potential barrier to the use of comparative judgement for summative teacher assessment is the relative inefficiency of the method compared to the marking procedure. Single-marking took approximately three hours, whereas comparative judging took approximately fourteen hours, albeit inherently moderated.

Comparative judgement assumes fast, intuitive decisions because, unlike marking, each test response is viewed several times across the judges. Moreover, evidence suggests that judges who spend longer making decisions are no more accurate than those who complete their judgements quickly (Jones et al., 2014; Pollitt, 2012). In the present study each test response was seen on average about twenty times across the five teachers, and on average each judgement took about 33 seconds, meaning each test response was eyeballed for about five and a half minutes. This is a long time considering that each test response took on average about 51 seconds to be marked by a single teacher.

A second, related issue is the number of judgements required for a reliable assessment outcome. Reliability can be increased to any desired level by increasing the number of judgments (quadrupling the number of judgements would halve the standard errors). Some have argued that efficiency savings can be made by considering carefully how test responses are selected and paired during the judging process (Bramley, 2007; Pollitt, 2012). In the present study pairings were selected entirely at random, and the number of views per test response ranged from 12 to 32. The software has subsequently been modified such that each test response is viewed an equal number of times, determined in advance.

Taken together, reducing the time taken for each judgement, and ensuring the test response selection procedure is leaner, offer a way forward for substantially increasing the efficiency of the comparative judgement approach. We note further that the relative inefficiency reported here is not the case for all comparative judgement studies (e.g. Jones et al., 2012; Kimbell, 2012; Bramley, 2007), and optimism about efficiency improvements seems warranted.

Moderation and standards monitoring. An alternative possible role for comparative judgement in teacher assessment is as a moderation tool. This might be particularly attractive in light of the efficiency challenges discussed above. Moreover, as some jurisdictions shift from an emphasis on centralised towards teacher-based national assessment (DfE, 2013; NCCA, 2009), and others question the reliability and validity of teacher assessment (Hume 2014; Jones, 2013), new approaches to moderation and standards monitoring are required (Brookhart, 2013).

Comparative judgement might be applied to a sample of student work from across different classes or different schools. For example, teachers might mark or rank their own student's work and supply a selected sample for a comparative judgement process involving all participating teachers or schools. In addition, samples might be included from previous cohorts to enable monitoring of standards over time as well as across different classes and schools. This latter possibility arises because comparative judgement can be applied to responses from different forms of tests (Jones et al., 2014). The outcome would allow inferences to be made about relative student performance across classes and schools, as well as the reliability of the teachers' assessments.

Integrating assessment and learning. The motivation for the study reported here was a teacher's response to a shift towards the use of teacher assessment for summative purposes in the Republic of Ireland. Commonly, the only outcomes of high-stakes assessments are grades and students are not provided with written or qualitative feedback on their work. This is consistent with the version of comparative judgement implemented here in which there is no scope of individualised feedback (although functionality for judge comments is included in other implementations; e.g. Seery, Canty & Phelan, 2012). Although level descriptors were provided to the students (see Appendix 2), the lack of individualised feedback, which is common to comparative judgement, presents a barrier to moving beyond summative assessment.

However, in the present study we took a small step towards evaluating the feasibility of integrating comparative judgement with teaching and learning by implementing a peer assessment exercise. The motivation for doing this was that involving students in the assessment of one another's work is reported to enhance learning (Strijbos & Sluijsmans, 2010), and comparative rather than absolute processes may be particularly beneficial (Pachur & Olsson, 2012). Peer assessment activities for learning can be particularly effective when students discuss and reflect on what constitutes high-quality work (Topping, 2009). Therefore a promising arrangement might be one in which students undertake comparative judgement activities, and consider and discuss how to improve their own understanding and communication in light of the relative efforts of others.

The multiple-regression analysis suggested that students, in contrast to the teachers, privileged factual recall over explanatory understanding when comparing pairs of responses. As such, tests based on an open item can be expected to present a significant challenge to students assessing one another's work. We might expect that students undertaking comparative judgement as part of structured lessons will both improve their understanding and communication of chemistry. This in turn could lead to better alignment between student and teacher judgements, thereby improving the validity of peer assessment outcomes.

Conclusion

The use of teacher judgement as part of summative assessment is a high-profile issue around the world, with some jurisdictions moving towards increasing its role and others reducing it. Here we have reported one teacher's response to the shift towards greater teacher assessment in Ireland using a comparative judgement approach. Our findings suggest that comparative judgement offers potential for summative assessment, although task design and efficiency need to be further investigated. In addition to summative assessment, comparative judgement also offers promise for moderating teacher assessment, and improving student learning through the use of peer assessment.

Acknowledgements

The research reported here was funded a Royal Society Shuttleworth Education Research Fellowship. We would like to thank the students and teachers involved in the reported work.

References

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research & Perspectives, 9*, 95-104.

ARG. (2006). *The Role of Teachers in the Assessment of Learning*. London: Assessment Reform Group.

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology, 85*, 347–356.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*, 215–232.

Black, P., & Wiliam, D. (2007). Large-scale assessment systems: Design principles drawn from international comparisons. *Measurement: Interdisciplinary Research and Perspectives, 5*, 1–53.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika, 39*, 324–345.

Bramley, T. (2007). Paired comparison methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 264–294). London: QCA.

Brookhart, S. M. (2013). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice, 20*, 69–90.

Choi, C. C. (1999). Public examinations in Hong Kong. *Assessment in Education: Principles, Policy & Practice, 6*, 405–417.

Crisp, V. (2013). Criteria, comparison and past experiences: how do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice, 20*, 127–144.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*, 438–481.

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*, 53–72.

Davies, D., Collier, C., & Howe, A. (2012). Assessing scientific and technological enquiry skills at age 11 using the e-scape system. *International Journal of Technology and Design Education, 22*, 247–263.

DfE. (2013). *Assessing Without Levels*. Online article retrieved 5.5.14 from http://tinyurl.com/omdpb3a

DES. (2012). *A Framework for Junior Cycle.* Dublin: Department of Education and Skills.

Donnelly, K. (2014). Teachers to vote on strike action over Junior Cert. *Independent.ie.* Retrieved 14 March 2014, from http://www.independent.ie/irish-news/teachers-to-vote-on-strike-action-over-junior-cert-30011709.html

Firth, D. (2003). Overcoming the reference category problem in the presentation of statistical models. Sociological Methodology, 33, 1–18. doi:10.1111/j.0081-1750.2003.t01-1-00125.x

Firth, D. (2005). Bradley-Terry models in R. *Journal of Statistical Software, 12*(1), 1–12.

Gill, T., & Bramley, T. (2013). How accurate are examiners' holistic judgements of script quality? *Assessment in Education: Principles, Policy & Practice, 20*, 308–324.

Harlen, W. (2004). *A Systematic Review of the Evidence of Reliability and Validity of Assessment by Teachers used for Summative Purposes*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher, 37*, 1–19.

Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology, 76*, 777–781.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*, 297–313.

Hume, A. (2014). Far North Queensland parents warned: Don't lean on NAPLAN. *The Cairns Post.* Retrieved 16.05.14 from http://tinyurl.com/mr7ddww

Husbands, C. T. (1976). Ideological bias in the marking of examinations: A method of testing for its presence and its implications. *Research in Education, 15*, 17–38.

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education, 28*, 91–105.

Jones, N. (2013). Thousands awarded wrong NCEA grades. *Otago Daily Times*. Retrieved 10.5.14 from http://tinyurl.com/lrnmu9e

Jones, I., & Alcock, L. (2013). Peer assessment without assessment criteria. *Studies in Higher Education,* 1–14. doi:10.1080/03075079.2013.821974

Jones, I. & Inglis, M. (submitted). The problem solving problem: Can comparative judgement help? *Educational Studies in Mathematics.*

Jones, I., Inglis, M., Gilmore, C., & Hodgen, J. (2013). Measuring conceptual understanding: The case of fractions. In A. M. Lindmeier & A. Heinze (Eds.), *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 113–120). Kiel, Germany: IGPME.

Jones, I., Swan, M., & Pollitt, A. (2014). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, 1–27. doi:10.1007/s10763-013-9497-6

Jones, I., & Wheadon, C. (submitted). Peer assessment using comparative and absolute judgement. *Educational Assessment.*

Kimbell, R. (2012). Evolving project e-scape for national assessment. *International Journal of Technology and Design Education, 22*, 135–155.

Laming, D. (1984). The relativity of 'absolute' judgements. *British Journal of Mathematical and Statistical Psychology, 37*, 152–183.

Lissitz, R. W. (2009). *The Concept of Validity: Revisions, New Directions, and Applications*. Charlotte, NC: IAP.

MacCann, R. G., & Stanley, G. (2010). Classification consistency when scores are converted to grades: examination marks versus moderated school assessments. *Assessment in Education: Principles, Policy & Practice, 17*, 255–272.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*, 13–23.

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

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment. *Assessment in Education: Principles, Policy & Practice, 21*, 205–220.

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

NCCA. (2009). *Innovation and Identity: Ideas for a New Junior Cycle*. Dublin: National Council for Curriculum and Assessment.

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology, 65*, 207–240.

Pollitt, A. (2004). Let's stop marking exams. Presented at *The International Association for Educational Assessment Conference*, Philadelphia.

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 19,* 281–300.

QCA. (2009). *Changes to GCSE Mathematics* (No. QCA/09/4159). London: Qualifications and Curriculum Authority.

Seery, N., Canty, D., & Phelan, P. (2012). The validity and value of peer assessment using Adaptive Comparative Judgement in design driven practical education. *International Journal of Technology and Design Education, 22*, 205–226.

Stanley, G., MacCann, R., Gardner, J., Reynolds, L., & Wild, I. (2009). *Review of Teacher Assessment: Evidence of What Works Best and Issues for Development*. Coventry: Qualifications and Curriculum Authority.

Strijbos, J.-W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments. *Learning and Instruction, 20*, 265–269.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286.

Tisi, J., Whitehouse, G., Maughan, S., & Burdett, N. (2013). *A Review of Literature on Marking Reliability Research* (No. Ofqual/13/5285). Coventry: Ofqual.

Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48*, 20–27.

Wiliam, D. (2001). Reliability, validity, and all that jazz. *Education 3-13: International Journal of Primary, Elementary and Early Years Education, 29*, 17–21.

Wilson, J., & Wright, C. R. (1993). The predictive validity of student self-evaluations, teachers' assessments, and grades for performance on the verbal reasoning and numerical ability scales of the differential aptitude test for a sample of secondary school students attending rural Appalachia schools. *Educational and Psychological Measurement, 53*, 259–270.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy & Practice, 17*, 59–75.