Assessment by Comparative Judgement: An Application to Secondary Statistics and English in New Zealand

Neil Marshall¹ · Kirsten Shaw¹ · Jodie Hunter² · Ian Jones³

Received: 4 November 2019 / Accepted: 31 March 2020 / Published online: 8 April 2020
© The Author(s) 2020

Abstract

There is growing interest in using comparative judgement to assess student work as an alternative to traditional marking. Comparative judgement requires no rubrics and is instead grounded in experts making pairwise judgements about the relative ‘quality’ of students’ work according to a high level criterion. The resulting decision data are fitted to a statistical model to produce a score for each student. Cited benefits of comparative judgement over traditional methods include increased reliability, validity and efficiency of assessment processes. We investigated whether such claims apply to summative statistics and English assessments in New Zealand. Experts comparatively judged students’ responses to two national assessment tasks, and the reliability and validity of the outcomes were explored using standard techniques. We present evidence that the comparative judgement process efficiently produced reliable and valid assessment outcomes. We consider the limitations of the study, and make suggestions for further research and potential applications.

Keywords Assessment · English · Statistics · Comparative judgement

Introduction

When a student’s assessment is assigned a grade we want that grade to be entirely dependent on the quality of the student’s work and not at all dependent on the biases and idiosyncrasies of whoever happened to assess it. Where we succeed we can say that the assessment is reliable, and where we fail we can say the assessment is unreliable (Berkowitz et al. 2000). Reliability in high-stakes educational assessment is

✉ Ian Jones
I.Jones@lboro.ac.uk

¹ New Zealand Qualifications Authority, Wellington, New Zealand

² Institute of Education, Massey University, Wellington, New Zealand

³ Mathematics Education Centre, Loughborough University, Loughborough, UK

particularly important where students' grades affect future educational and employment prospects.

One way to ensure reliability is to use so-called objective tests in which answers are unambiguously right or wrong (Meadows and Billington 2005). Examples are an arithmetic test in mathematics, a spelling test in languages, or a multiple-choice test in any subject. Objective tests also have the advantage of being quick and inexpensive to score and grade because the task can be automated (Sangwin 2013) and in practice often is (Alomran and Chia 2018). For these reasons—reliability and efficiency—objective tests are common in education systems around the world (Black et al. 2012).

However, objective tests are far from universal because they risk delivering reliability at the cost of validity (Assessment Research Group 2009; Wiliam 2001). Educationalists have argued that there is more to doing and understanding mathematics or languages than answering a series of closed questions. We typically want evidence that students can make connections between learned ideas, apply their understanding to novel contexts, construct arguments, demonstrate chains of reasoning, and so on (Newton and Shaw 2014; Wiliam 2010). Such achievements are not readily assessed using objective tests but instead require students to undertake a sustained performance that reflects the discourse and cultural norms of the subject (Jones et al. 2014). In mathematics this might be a student report in which mathematical ideas are creatively applied to a problem situation; in languages this might be a student report in which literary texts are critiqued. Scholars have argued that such assessments are more valid and fit for purpose than objective tests (e.g. Baird et al. 2017). However, assessing performance-based assessments is subjective and therefore tends to be less reliable than objective tests, in part because outcomes reflect to some extent the biases and idiosyncrasies of individual assessors (Murphy 1982; Newton 1996).

As such there exists a tension between ensuring an assessment is reliable and ensuring that it is valid (Wiliam 2001; Pollitt 2012). To negotiate this tension education systems commonly make use of three techniques. First, many tests are populated with short questions, typically worth several marks, which might be seen as a compromise between objective-style and performance style questions (Hipkins et al. 2016). Second, many exams and assessments are designed with a rubric or marking scheme that assessors use as the basis for awarding marks. Rubrics typically attempt to anticipate the full range of student answers, although in practice professional judgement is often required when matching a script to a rubric (Suto and Nadas 2009). The third technique is for a sample of marked tests to be moderated, or re-marked, by an independent assessor.

However these techniques produce imperfect outcomes. Short questions have been criticised for being more similar to objective questions and contributing to a tick-box approach to assessment (Assessment Research Group 2009; Black et al. 2012). Where questions are more substantive, rubrics become so open to interpretation that reliable assessment outcomes are not achieved (Meadows and Billington 2005; Murphy 1982; Newton 1996). Moderation exercises can only be applied to a sample of student scripts or the marking workload is doubled, and so commonly a student receives a grade without their particular script having been moderated.

In this paper we provide evidence in support of the reliability and validity of an alternative approach to producing student assessment outcomes, called comparative judgement (Pollitt 2012). We describe the origins and mechanics of comparative judgement and argue that there are educational assessment contexts in which it can deliver acceptable reliability and validity. We then present two studies in which we evaluated the application of comparative judgement to assess standard secondary school statistics and English assignments in New Zealand.

Comparative Judgement

Comparative judgement is a long-established research method that originates in the academic discipline of psychophysics. Work by the American psychologist Thurstone (1927) established that human beings are consistent with one another when asked to compare one object with another, but are inconsistent when asked to judge the absolute location of an object on a scale. For example, Thurstone found that participants were consistent when asked which of two weights is heavier and inconsistent when asked to judge the weight of a single object. Thurstone (1954) later used comparative judgement to construct scales of difficult-to-specify constructs such as beauty and social attitudes.

The recent two decades has seen the growth of research and practice in applying comparative judgement to educational assessment. The internet has enabled larger scale of usage than is possible in traditional laboratory settings because student work can be digitised and presented to examiners remotely and efficiently (Pollitt 2012). A key motivation for considering comparative judgement rather than traditional assessment methods are claims that comparative judgement is more reliable than marking for the case of open-ended assessments (Jones and Alcock 2014; Steedle and Ferrara 2016; Verhavert et al. 2019).

Once the judging is complete the examiners' binary decisions are fitted to a statistical model to produce a unique score for each script (Bramley 2007). It is not necessary to compare every script with every other script; the literature suggests that to assess n scripts then a total $10n$ to $37n$ judgements are required (Jones and Alcock 2014; Verhavert et al. 2019), depending on the nature of the tests and the expertise of the assessors, to produce a reliable scale. The reliability and validity of outcomes can be estimated using the techniques described in the methods section. Good reliability and validity for assessment outcomes have been reported in a variety of contexts including mathematical problem solving (Jones et al. 2014; Jones and Inglis 2015) and essays (Heldsinger and Humphry 2010; Steedle and Ferrara 2016; van Daal et al. 2019).

There are two key reasons that comparative judgement outcomes can be more reliable than traditional marking when assessing student responses to open assessment tasks. First is people's capacity to compare two objects against one another more consistently than they can rate an object in isolation, as discussed above. Second is that assessor bias has no scope for expression when making binary decisions rather than assigning marks (Pollitt 2012). For example, a harsh assessor is likely to assign a lower mark than a lenient assessor. However when comparing two pieces

of work both assessors must simply decide which piece of work is the better, and so they produce the same outcome.

The Research

We contribute to literature on using comparative judgement for the case of secondary assessments in statistics (Study 1) and English (Study 2) in New Zealand. Our focus was on whether the good reliabilities reported in the literature could be replicated for the case of secondary assessments in New Zealand. We also explored validity, and were able to do so in depth for the case of Study 1 where relevant data were available. Finally, we sought to understand the judges' experience of the comparative judgement process using a survey.

Study 1: Statistics

Materials

The assessment task used was for NCEA Level 2 AS91264 'Use Statistical methods to make an inference' and is shown in Appendix 1. The task requires students to take a sample from a population and make an inference that compares the position of the median of two population groups.

Participants

The participants were 134 students in Year 12 (ages 16 and 17) from a single school in New Zealand. The school was decile 8 had a roll of about 1400 students with 60% of students of NZ European ethnicity, 14% of students with Maori ethnicity and 5% of students of Pasifika ethnicity. The school was recruited by contacting the Head of Department of mathematics through a contact of one of the authors and individual teachers were invited to participate on a voluntary basis. The study followed NZQA ethical practices for conducting research and guidelines for assessment conditions of entry.

Method

The students completed the assessment task as a summative assessment of a unit of work that was taught over a period of 6–8 weeks. Students completed the assessment electronically in Google Drive over four days. They could work on the assessment task in and out of normal lesson time. 21 scripts were incomplete and were excluded from further analysis, leaving a total 113 participant scripts. The anonymised student scripts were uploaded to the online comparative judgement engine nomoremarking.com for assessment. In addition six boundary scripts,

produced by NZQA to exemplify levels of achievement (High Not Achieved, Low Achieved, High Achieved, Low Merit, High Merit, Low Excellence), were also uploaded. Therefore the total number of participant and boundary scripts to be judged was 119.

The scripts were judged by 21 judges: eight judges were teachers from the same school as the participants; eight were pre-service teachers undertaking a practicum at a range of New Zealand schools; five were NZQA employees who work as National Assessment Moderators, and one of whom had previously worked as a subject advisor. The project included two initial sessions of professional development for the judges prior to the assessment tasks being undertaken. The first session, led by the NZQA moderator, supported assessors to become fully familiar with the assessment activity and the corresponding achievement standard. The second session supported assessors to become fluent in using the comparative judgement online process. The judges then undertook the assessment procedure using this online comparative judgement process (nomoremarking.com). The judges were allocated 99 pairwise comparisons each. Eighteen judges completed all 99 comparisons and the other three completed 16, 50 and 63 each. This resulted in 1911 judgements in total with each script receiving between 31 and 36 judgements each with a mode of 32 judgements.

We also collected participants' Grade Point Averages (GPAs) for mathematics and English from their Year 11 results, and these data were available for 99 of the participants.¹ It was not possible to calculate a GPA for the other 14 students because they arrived at the school from overseas, either as the result of family immigration or as exchange students.

Analysis and Results

Comparative Judgement Scores

The judgement decisions were fitted to the Bradley-Terry model to produce a unique score for each student script (Pollitt 2012). The scores had a mean of 0.0 and a standard deviation of 1.7. The distribution of scores is shown in Fig. 1.

Reliability

Reliability was explored using three techniques (Jones and Alcock 2014). First, we calculated the Scale Separation Reliability (SSR), which gives an overall sense of the internal consistency of the outcomes considered analogous to Cronbach's alpha

¹ The GPA is calculated from assessed achievement standards. The GPA is a weighted score, ranging from 0 to 400, calculated for each student as $100 \times \frac{\sum_{i=1}^n C_i G_i}{\sum_{i=1}^n C_i}$, where n is the number of achievement standards entered, and for each achievement standard C_i is the number of credits (ranges from 2 to 5) and G_i is one of four grades awarded (not achieved = 0 points, achieved = 2 points, merit = 3 points, excellence = 4 points).

Fig. 1 Comparative judgement scores for statistics scripts

(Pollitt 2012). A high value, $SSR \geq 0.7$, suggests good internal consistency and this was the case here, $SSR = 0.90$.

Second, we estimated inter-rater reliability using the split-halves technique as described in Bisson et al. (2016). The judges were randomised into two groups and each group's decisions refitted to the Bradley Terry model to produce two sets of scores for the scripts. The Pearson product–moment correlation coefficient is calculated between the two sets of scores, and this process is repeated 100 times and the median correlation coefficient taken as an estimate of inter-rater reliability. The inter-rater reliability in this case was strong, $r = 0.81$.

Third, we calculated an infit statistic for every judge and compared these statistics to a threshold value (two standard deviations above the mean of the infit statistics) to check for any 'misfitting' judges (Pollitt 2012). An infit statistic above the threshold value suggests that judge's decisions were not consistent with the other judges' decisions. There were no misfitting judges bar one who was marginally above the threshold value. We recalculated comparative judgement scores with the misfitting judge removed and found the correlation with the original scores (with the misfitting judge included) to be perfect, $r = 1$. Therefore the misfitting judge had no effect on final outcomes and was included in the remainder of the analysis.

Taken together, these analyses suggest that the scores were reliable; that is we would expect the same distribution of scores had an independent sample of experts conducted the judging.

Validity

The validity of assessment outcomes is more difficult to establish than reliability (William 2001). This difficulty is in part because whereas reliability is an inherent property of assessment outcomes, validity tends to require comparison with external data such as independent achievement scores and expert judgement (Newton and

Fig. 2 Scatter plots of the relationship between comparative judgement scores and GPAs

Shaw 2014) and needs to take into account the intended purpose and interpretation of test scores (Kane 2013). In terms of the purposes of test scores, a limitation of much assessment research is that tests are often administered in contexts that are low- or zero-stakes from the perspective of the student participants. This threat to ecological validity was overcome in the present study which made use of responses gathered as part of students' usual summative assessment activities.

For Study 1, we investigated validity using three techniques as described here. First, we explored criterion validity (Newton and Shaw 2014) by scrutinising the positioning of the boundary scripts in the final rank order. The boundary scripts represented grades at six levels (High Not Achieved, Low Achieved, High Achieved, Low Merit, High Merit, Low Excellence) and were determined using traditional marking methods. A valid comparative judgement assessment would be expected to position the boundary scripts in the correct order, from High Not Achieved to Low Excellence, and this was the case as shown in Fig. 1.

Second, we investigated convergent and divergent validity (Newton and Shaw 2014) using Mathematics and English GPAs that were available for 99 of the 113 students who completed the statistics test from their Year 11 results. The validity of the comparative judgement scores would be supported by evidence of convergence with mathematics and divergence with English GPAs. To investigate this we regressed mathematics and English GPAs onto comparative judgement scores. The regression explained 26.9% of the variance in comparative judgement scores, $F(2,96) = 17.67$, $p < 0.001$, $\eta^2 = 0.269$. Mathematics GPA was a significant predictor of comparative judgement scores, $b = 0.052$, $p = 0.002$, and English GPA was also a significant predictor, $b = 0.044$, $p = 0.001$. This was not expected because previous research conducted in the UK has consistently found that mathematics, but not English, attainment is a significant predictor mathematics comparative judgement scores (Jones et al. 2013; Jones and Wheadon 2015; Jones and Karadeniz 2016). To understand the relationship between the comparative judgement scores and GPAs

Fig. 3 Comparative judgement scores and teacher grades for the statistics scripts

further we investigated the correlations. The comparative judgement scores correlated moderately with mathematics GPAs, $\rho=0.450$, and moderately with English GPAs, $\rho=0.468$, as shown in Fig. 2, and these correlations were not significantly different (Steiger, 1980), $z = -0.19$, $p=0.851$. Interestingly, the correlation between mathematics and English GPAs was similar, $\rho=0.390$. Therefore, the correlations between the comparative judgement outcomes and mathematics and English GPA are in line with what we might generally expect from different assessments.

Third, to further explore criterion validity, teachers at the school (who were also part of the comparative judgement process) awarded grades (Not Achieved, Achieved, Merit, Excellence) to the scripts using the traditional marking and internal moderation process that was familiar to them. Following usual protocol, each teacher awarded a grade to the students in his/her class and the name of the student was visible to the teacher. A small sample of scripts from each class were check marked by the Head of Department for internal moderation processes. The teachers' grades were further moderated by National Assessment Moderator employed by NZQA. The moderator reviewed each script and confirmed, or not, the teacher judgement. Where there was disagreement the moderator judgement took precedence. Teacher grades were available for 94 of the participants.

A valid comparative judgement assessment would be expected to produce increasing scores for those scripts graded at Not Achieved through to those graded at Excellence. Indeed the correlation between comparative judgement scores and grades was moderate, $\rho=0.740$, as shown in Fig. 3. A one-way ANOVA was conducted to compare the comparative judgement scores of scripts in each of the four grades. There was a significant difference of scores at each grade, $F(3, 96)=44.81$, $p<0.001$. Using a Bonferroni-adjusted alpha level of 0.017 there was a significant difference between scripts graded by teachers as Not Achieved ($M = -1.98$, $SD=1.17$) and Achieved, ($M = -0.14$, $SD=0.86$), $t(32.49) = -5.69$, $p<0.001$. There was a marginal difference between scripts graded as Achieved and Merit ($M=0.64$, $SD=1.12$), $t(47.57) = -2.44$, $p=0.019$. There was a significant difference between

scripts graded as Merit and Excellence ($M=1.37$, $SD=0.92$), $t(50.55)=-3.07$, $p=0.003$.

Taken together these three analyses support the validity of using comparative judgement to assess the statistics task.

Study 2: English

Materials

The assessment task used was NCEA Level 1 AS90852 ‘Explain significant connection(s) across texts, using supporting evidence’, as shown in Appendix 2. The task required students to write a report in which they identify a connection across four literature texts. Three of the texts were teacher-selected and one was student-selected. Students identify examples of their chosen connection from each of the literature texts and then explain them. The assessment activity was open book and completed online within a specified timeline.

Participants

The participants were 253 students in Year 11 (ages 15 and 16) from a single school in New Zealand. The school was not that in Study 1. It was decile 8 with a roll of about 1300 students with 55% of students of NZ European ethnicity, 27% of students with Maori ethnicity and 7% of students of Pasifika ethnicity. Recruitment and ethical practices were the same as for Study 1.

Method

The method was broadly the same as Study 1, although unlike Study 1 boundary scripts were not available and none were included in the judging pot. The 253 scripts were judged by 17 judges: ten judges were teachers from the same school as the participants; seven were NZQA employees who work as National Assessment moderators. There were two initial sessions of professional development for the judges, led by the NZQA moderator. The first session supported assessors to become fully familiar with the assessment activity and the corresponding Achievement Standard, and the second session showed assessors how to comparatively judge scripts online. The judges were allocated 119 pairwise comparisons each; 15 judges completed all 119 comparisons and the other two completed 20 and 51 each. This resulted in 1856 judgements in total with each script receiving between 13 and 19 judgements each with a mode of 15 judgements.

Fig. 4 Comparative judgement scores for English scripts

Analysis and Results

Comparative Judgement Scores

Fitting the decision data to the Bradley Terry model produced comparative judgement scores with mean 0.0 and standard deviation 2.4, and the distribution is shown in Fig. 4. The boundaries in Fig. 4 are between two scripts whereas in Fig. 1 the boundaries were scripts within the rank order. This difference across the two studies arose because boundary scripts were not available for inclusion in the judging pot for English and consequently the grade boundaries were derived quite differently: six scripts were selected using teacher knowledge of the students grounded in classroom formative assessment activities. Two of these scripts were expected to be on the boundary of Not Achieved and Achieved, another two on the boundary of Achieved and Merit and a further two on the boundary of Merit and Excellence. In every case these were confirmed as boundary scripts, with one above the boundary and one below. The six scripts were also graded by the Head of Department and the National Moderator, and their position noted in the comparative judgement ranking. A further two scripts either side of these two in the ranking were graded in the same way. Thus at each boundary six scripts were graded by the traditional method and the boundary within the comparative judgement ranking established. For each boundary this proved to be a reliable process: the scripts either side of the initial two scripts were graded as weaker and stronger than the first two.

Reliability

The internal consistency of the comparative judgement scores was strong, $SSR=0.85$ and the split-halves inter-rater reliability was moderate, $r=0.72$. The reliability of comparative judgement scores is related to the number of judgements

Fig. 5 Comparative judgement scores and two sets of independent examiner grades for the English scripts

(Verhavert et al. 2019), and the lower values here compared to Study 1 ($SSR=0.90$, $r=0.81$) would be expected because there were fewer judgements per script. There were no misfitting judges for the case of the English assessment. Taken together these analyses provide some support for the reliability of the assessment outcomes.

Validity

Unlike Study 1, GPAs were not available for the participating students in Study 2. Instead, two experienced teachers from schools not involved in the comparative judgement study independently assigned grades (Not Achieved, Achieved, Merit, Excellence) to a random sample of 50 scripts using traditional marking processes. The agreement between the examiners' grades was acceptable, $\kappa=0.76$.

To evaluate criterion validity we investigated whether the comparative judgement scores increased from scripts graded at Not Achieved through to those graded at Excellence. This was confirmed as shown in Fig. 5. A one-way ANOVA was conducted to compare the comparative judgement scores of scripts in each of the four grades for the case of the first examiner. There was a significant difference of scores at each grade, $F(3, 46)=21.34$, $p<0.001$. Using a Bonferroni-adjusted alpha level of 0.017 there was a significant difference between scripts graded as Not Achieved ($M=-2.14$, $SD=1.20$) and Achieved, ($M=0.23$, $SD=1.93$), $t(27.16)=-4.19$, $p<0.001$. There was no significant difference between scripts graded as Achieved and Merit ($M=0.60$, $SD=1.59$), $t(15.94)=-0.62$, $p=0.545$. There was a significant difference between scripts graded as Merit and Excellence ($M=3.02$, $SD=1.46$), $t(16.99)=-3.45$, $p=0.003$. These results were replicated for the grades of the second examiner, although the difference between scripts graded as Merit and Excellence was marginal with a Bonferroni-adjusted alpha level of 0.017, $t(17.30)=-2.56$, $p=0.020$.

Table 1 Summary of judges' responses to the demographic and quantitative survey questions

	Statistics	English
Totals	14	12
Demographic questions		
Gender		
Male	4	2
Female	10	10
Age		
26–35 years	1	2
36–45 years	6	3
> 45 years	7	7
Role		
Teacher	9	8
NZQA moderator	7	7
Learning provider	1	0
Teaching experience		
5 – 9 years	1	2
> 9 years	8	7
NA	5	2
Quantitative questions		
1. Registering on the website to make judgements was (1 = very easy, 5 = very difficult)	2.2 (1.5)	2.4 (1.6)
2. The instructions to get started were (1 = not helpful, 5 = very helpful)	4.1 (0.9)	3.4 (1.3)
3. Comparing two scripts on the website was (1 = very easy, 5 = very difficult)	3.1 (1.4)	3.1 (0.9)
4. Comparing documents with several pages was (1 = very easy, 5 = very difficult)	3.1 (0.9)	3.1 (0.8)
5. Comparing different response formats (eg Word v PPT) was (1 = very easy, 5 = very difficult)	2.7 (0.8)	2.8 (0.7)
6. When making judgements, I decided using intuition (1 = never, 5 = all the time)	2.3 (1.0)	3.4 (0.7)
7. I decided by carefully matching responses to the standard (1 = never, 5 = all the time)	4.4 (0.7)	3.3 (0.9)
8. I forget to log off when I took a break from judging (1 = never, 5 = often)	3.1 (1.6)	2.4 (1.3)

Figures shown for the quantitative questions are mean (standard deviation)

Survey

A survey was conducted to further investigate the validity, from the perspective of the judges, of the assessment process across both studies. The survey was designed to collect quantitative data on judges' demographics and experience of conducting comparative judgement, including how they made decisions. The survey also

collected qualitative data through two open-text questions that asked for any other feedback about the marking experience using comparative judgement. All judges involved in the project were invited to complete the survey and 26 or 68% (14 for statistics, 12 for English) did so. The quantitative questions and responses are shown in Table 1, and the open-text questions and responses are shown in Appendix 3.

Table 1 shows that the demographic and quantitative questions received broadly similar patterns of responses from judges in both studies. Overall the respondents reported that mechanics of the process were relatively easy (Q1 and Q2), and that making judgement decisions was moderately difficult (Q3 to Q5).

An interesting difference between the studies is that the statistics judges reported being less likely to use intuition (Q6) than the English judges, Mann–Whitney $U=137.0$, $p=0.005$. Conversely, the statistics judges reported being more likely to match responses to the standard (Q7) than the English judges, Mann–Whitney $U=30.0$, $p=0.004$. This difference might reflect our expectations for each subject; assessing English involves more subjective judgement whereas assessing statistics is a more objective activity of judging correctness. Indeed, evidence of mathematics teachers and examiners mentally ‘converting’ scripts to grades and then conducting a comparative judgement between those grades has been reported previously (Jones et al. 2014). Nevertheless, judges in both studies commonly converted scripts to grades in their heads and then compared these estimated grades, as reflected in Q7 as well as in eight of the open-text responses. For example, one English judge wrote “comparing two pieces of work that both were at, for example, Merit seemed pointless as we do not differentiate between high and low Merit”.

Converting to grades and then comparing those grades is not in the intended spirit of comparative judgment because mental grading involves comparing each script against an absolute standard rather than directly comparing one script against another. It is the authors’ experience that entering the spirit of comparative judgement is a difficult transition for some teachers who are used to traditional marking methods. One statistics judge wrote “sometimes found it hard to break from tradition and I wanted to grade [the scripts]”. The mental application of grades to scripts prior to comparison left some judges of the view that comparative judgement is an inappropriate or ineffective method of assessment, a sentiment that was explicit or implicit in five of the open-text responses. This may have impacted on the reliability and validity of the outcomes.

A related concern, evident in four of the open-text responses, was that relative judgements cannot be used to grade scripts. As one English judge wrote, comparative judgement consists of “ranking texts that should be assessed against the standard; so ranking could be completely irrelevant (e.g. the highest may still not actually achieve)”. This reflects a common intuition that comparative judgement exclusively produces norm rather than criterion-referenced outcomes. However this is not the case: scores can be used to rank or to grade whether they result from comparative

judgement or traditional marking, as we have demonstrated using different techniques in both the studies reported here.

Eight of the open-text responses commented on the time required to do judging. One respondent stated that it was a “time saver”, another that they were “fixated” about how long each judgement was taking, and the remaining six that it was time consuming, presumably in comparison to traditional marking. This is interesting and perhaps surprising in the light of claims that comparative judgement is more efficient in terms of assessor time (Steedle and Ferrara 2016). We can roughly estimate total judging time because the *nomoremarking* website records the amount of time taken for each judgement. These time data are an overestimate because the timer continues when judges take a break, and we know from Q8 that the judges did not always log off when taking breaks. However we can take the mean of the judges’ median judging time as an estimate and multiply this by 10, which is the typical number of judgements per script required for internally consistent outcomes² (Jones and Alcock 2014; Jones and Inglis 2015). We then halve this value because every pairwise judgement involves two scripts. This rough calculation suggests that for both statistics and English each script takes about 7 min to assess on average. Anecdotally, statistics teachers inform us that this is approximately half the time taken to grade by a traditional marking method. For the case of English, the two examiners who marked a random sample scripts each reported this taking around 6 h to complete 50 scripts, thereby matching the estimate of 7 min per script for comparative judgement. In addition, comparative judgement produces inherently moderated outcomes, because every script has been seen by several assessors, whereas the estimate for traditional marking does not include time for moderation.

Discussion

In the studies reported here we explored the application of comparative judgement to the assessment of statistics and English in New Zealand. Across both applications we reported that the comparative judgement procedure used here produced reliable outcomes; i.e., outcomes were largely independent of the particular individuals who conducted the comparative judging online. We also found evidence supporting the validity of the approach in each discipline, by comparing comparative judgement outcomes with achievement data and independently marked scripts. Further evidence for validity was provided by responses to a survey from 68% of the judges across the two studies. This evidence also revealed differences between the two subjects, with statistics judges reporting greater matching of scripts to Achievement Standards, and English judges reporting greater use of intuition. Finally, we considered the efficiency of the comparative judgement process relative to traditional marking methods, and concluded that comparative judgement produces moderated outcomes in the same or less time than marking produces unmoderated outcomes.

² In the studies reported here we collected > 10 judgements per script to enable us to estimate inter-rater reliability. However this is not necessary and is not generally done in operational applications of comparative judgement.

Therefore we conclude that the comparative judgement technique as applied here efficiently produced moderated assessment outcomes that were reliable and valid. This conclusion applied to the context of particular secondary students' responses to standard statistics and English assessment tasks, and contributes this context to the growing literature that evaluates the application of comparative judgement to a varied ranged educational assessment activities. In the remainder of the discussion we consider limitations of the study and issues that might arise were a comparative judgement-based approach to assessment to be applied more generally.

The key limitation of the study design is common to assessment research: it is relatively straightforward to evaluate the reliability of an assessment, at least in the sense of the extent to which the final scores were independent of the particular assessors who produced them as was the focus here. However, evaluating validity is a less straightforward matter and requires the triangulation of quantitative and qualitative data to construct a validation argument that accounts for the intended purposes of the scores (Kane 2013; Newton and Shaw 2014). We attempted this here with the data available to us including prior achievement data, consideration of grade boundaries, independent assessment of responses and a survey of assessors. However, validation can never be fully demonstrated or completed. Future research might harness further validation methods including the use of independent measures of the assessed construct (e.g. Bisson et al. 2016), standardised measures of potential covariates (e.g. Jones et al. 2019), content analysis (Hunter and Jones 2018), comparison of expert and non-expert judge outcomes (Jones and Alcock 2014), and other methods.

There are also limitations to the study design and potential applications of comparative judgement that are particular to the standard assessments considered in this paper. For example, each student response was assigned one of four grades, whereas comparative judgement results in a unique score for each script. As such comparative judgement scores need then be translated into grades, as was done here through the use of boundary scripts (statistics) or boundary setting (English). A consequence translating scores into grades is that some pairwise judgements contribute little to the final outcome; as we reported above one judge commented that this is little point choosing one Merit script as over another Merit script. Comparative judgement might therefore be too precise a measure of achievement for certain NCEA level 1 and 2 tasks; or to put it another way comparative judgement might be *too* reliable for contexts where grades are the only required outcome. However, reducing the reliability through, say, using fewer pairwise judgements to construct scores might not be viable because it would risk blurring the boundaries and thereby increase the number of students who receive the 'wrong' grade. To investigate this we conducted a post-hoc analysis on the statistics data using only one third of the available judgements, resulting in a set of comparative judgement scores with lowered reliability ($SSR=0.74$ rather than the $SSR=0.90$ reported in Study 1). We then assigned one of four grades (Not Achieved, Achieved, Merit, Excellence) to each script with reference to the boundary grades for both the original and the lowered-reliability set of scores. We found that 44 (39%) of students would be assigned a different grade if the lowered-reliability scores were used in place of the original Study 1 scores. Therefore while the precision of comparative

judgement might be wasted on discriminating scripts that are well within grade boundaries, it is required to ensure adequate discrimination near grade boundaries. Notwithstanding comparative judgement was as or more efficient than traditional marking in both studies, there seems nevertheless to be an inherent inefficiency when the method is used in contexts where scores are translated into grades.

A second possible weakness of the results reported here is the use of a small number of boundary scripts to set grade cut offs.³ Boundary scripts have been used in other comparative judgement-based studies in mathematics (e.g. Jones and Sirl 2017) and in English (e.g. Heldsinger and Humphry 2010) but here the grade boundaries reflect quite specific and detailed NCEA standards, and we have assumed that judges are making their comparisons on a dimension that is appropriate to the standard. Our interview data provide some support that judges are mindful of the standards when judging, but more systematic methods to boundary setting could be employed. For example, Heldsinger and Humphry (2010) reported a systematic comparative judgement-based method for generating boundary scripts for standard setting in primary English in Australia. Such a method might be adapted to present NCEA practice by using a standard-setting procedure in which a sample of scripts are read and holistically compared with the standard for a grade.

Alternatively, comparative judgement methods might be better suited to contexts in which the task and assessment criterion are designed with pairwise judgement by subject experts in mind from the start (Jones and Inglis 2015). That is, educational assessment tasks should be designed with the method of converting student responses into scores, whether by multiple-choice, rubrics and marking, or indeed comparative judgement, considered from the outset. For the case of comparative judgement this might lead to more open and less predictable assessment tasks than those currently offered. Given the consensus that a rich diet of assessment is necessary to maximise validity and equity in educational outcomes (e.g. Wiliam 2010), comparative judgement might have an important part to play in the national education systems of the future.

Acknowledgements The authors are grateful to the schools and students who made this study possible.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix 1: Assessment Task NCEA Level 1 AS91264 'Use Statistical Methods to Make an Inference' Used in Study 1

Students received a sample of a dataset containing 800 rows that we are unable to reproduce here.

³ We are grateful to a reviewer of an early draft of this manuscript for this observation.

Student Instructions

Introduction

The dataset contains data that was collected from the student management system of High School for Year 11 and 12 students attending school in 2015.

The data was extracted from the student management system KAMAR by the school's data analyst Stefan Dale. Only Year 11 and 12 year group data were downloaded. These data were not cleaned and should be treated as raw data. No Year 11 or 12 students were harmed in the collection of these data.

Variable	Description
Gender	Male/female: M/F
Year level	Year level of students, either year 11 or year 12
Periods—80%	Attendance rates: ABV—over 80% of attendance or BELOW—under 80% of attendance
NZQA—Credits Current Year—Attempted	Number of credits attempted during the year—results could be N/A/M/E. Not Achieved results mean no credits are received
NZQA—Credits Current Year—Not Achieved	Attempted credits—Total credits → these credits are not counted towards total credits
NZQA—Credits Current Year—Achieved	Number of Achieved credits obtained during the year
NZQA—Credits Current Year—Merit	Number of Merit credits obtained during the year
NZQA—Credits Current Year—Excellence	Number of Excellence credits obtained during the year
NZQA—Credits Current Year—Total	Total number of Achieved, Merit & Excellence credits obtained during the year
Periods—Percentage (Year)	Percentage of periods of class attended during the year (see Periods—80%)

Task

Carry out a comparative statistical investigation using the data from the 2.9 Absences Year 11 and 12 2015 data set and prepare a report on your investigation.

You will need to complete all of the elements of the statistical enquiry cycle:

- Pose an appropriate comparative investigative question that can be answered using the data provided in the 2.9 Absences Year 11 and 12 2015 data set.
- Record what you think the answer to your question may be.
- Select random samples to use to answer your investigative question. You need to consider your sampling method and your sample size.

You will assume the data set is the Population. You must select a random sample in this assessment to demonstrate your understanding of sample to population inference

- Select and use appropriate displays and measures.
- Discuss sample distributions by comparing their features.
- Discuss sampling variability, including the variability of estimates.
- Make an inference.
- Conclude your investigation by answering your investigative question.

Prepare a report that includes:

- Introduction stating your comparative investigative question, the purpose of your investigation and what you predict the result will be.
- Method you used to select samples and collect data.
- Results and analysis of your data.
- Discussion of your findings and conclusion.

The quality of your discussion and reasoning and how well you link the context to different stages of the statistical enquiry cycle will determine the overall grade.

Appendix 2: Assessment Task NCEA Level 1 AS90852 'Explain Significant Connection(s) Across Texts, Using Supporting Evidence' Used in Study 2.**Student Instructions****Introduction**

This assessment activity requires you to write a report (of at least 350 words) in which you explain significant thematic connections across at least four selected texts. This will take place during the year's English programme.

You will have the opportunity to receive feedback, edit, revise, and polish your work before assessment judgements are made.

You can read texts, collect information, and develop ideas for the assessment activity both in- and out-of-class time.

You will be assessed on how you develop and support your ideas, and on the originality of your thinking, insights, or interpretation.

Preparatory Tasks**Choose a Theme and Texts**

Choose a theme. Your theme may arise from books that you have read, films you have seen, favourite songs, television programmes, or other class work.

Choose your texts. At least one text must be one that you have read independently. Your teacher will guide you in your choice of an independent text.

Keep a Record of the Texts as You Read

Draw up a record sheet and, over the course of the year, record some of the ways in which texts you read or view are connected to your chosen theme. See Resource A for examples of the kinds of statements you could make.

Task

Write a report of at least 350 words.

Begin with an introduction that identifies your texts and the connections between them.

Explain how each individual text is connected to the theme and/or the other texts. You may identify more than one connection across some of the texts.

Refer to specific, relevant details from each text that illustrate the connection across your texts.

Make clear points that develop understandings about the connections being identified.

Show some insight or originality in your thought or interpretation.

Appendix 3: Open-Text Responses from the Survey

#	How else did you make judgements? What, other than intuition and the standard influenced your decision?	We would be grateful for any other feedback that you have about your experience with using nomoremarking for NCEA internally assessed achievement standards
1	Key pieces of evidence showing partial understanding of the achievement criteria when both scripts were at the same grade	Some of the scripts being compared were both not Achieved which made it difficult assigning the 'better' script. Some of the scripts seemed to have students overwriting exemplars or other tasks. It was difficult to make assign the 'better' script when there were inconsistencies in the evidence provided. For example having some contradictory conclusions or changing the context to a totally different investigation within the discussion
2	By looking at what is required for the std—qty of texts, link to connection	It was harder to make the judgements than I thought it was going to be
3	NA	I found the 'timer' a bit distracting or rather I became a bit 'fixated' on how much time I was taking per answer
4	Through looking at key parts	It was interesting but the whole process has made for more work and a much longer lag time for the grading and recording

#	How else did you make judgements? What, other than intuition and the standard influenced your decision?	We would be grateful for any other feedback that you have about your experience with using nomoremarking for NCEA internally assessed achievement standards
5	Professional knowledge	Comparing two pieces of work that both were at, for example, merit seemed pointless as we do not differentiate between high and low merit. Secondly, this is ranking texts that should be assessed against the standard- so ranking could be completely irrelevant (eg the highest may still not actually achieve)
6	It really was simply comparing one to the other using my professional knowledge and awareness of the standard	I really liked it. I wonder how it would work when actually assigning a NAME grade- for now I saw/see it working for 'which one is better'- how does this translate to NAME?
7	Looking for key components that signify a fail grade. Used my memory (the paper had appeared before for comparison)	It is very good that you don't have to do it all in one sitting
8	Previous experience	It was not as effective as I thought it would be
9	By having to accept that some components were more important than others which is not what assessing this standard is about	This standard cannot deem one Not Achieved script as better than another Not Achieved script as all components of the statistical cycle need to be addressed. This standard cannot be assessed in the way suggested
10	Careful reading and reflexing in my head	I think it was a lot more work than just marking your own although for professionally learning it was good to read responses from other classes
11	Previous judgements	this took longer than anticipated
12	Knowing the students in my class	It was time consuming
13	Accidentally clicking the left/right button; professional judgement	Sometimes found it hard to break from tradition & wanted to grade then particularly when both scripts would be marked were N—which is the better N? Would like to be able to zoom the scripts to make it easier to read—on my mac the surrounding txt zoomed but not scripts
14	The evidence of 'going beyond the texts' and ability to make some sort of critical response to and evaluation of the ideas in the texts	I didn't have a clear idea of the concept as I was working through the comparative marking—not sure of how it worked and how it related to making grade boundary judgements etc.
15	Knowledge of texts, and identifying areas where students might show more original ideas rather than taught ideas	It was quite a difficult task at times. The number of judgments was a lot, and I wonder if it is quicker to mark own class with student work that I'm already familiar with. Seemed a little burdensome, and could be very hard making judgments between such different work
16	Quantity and quality	Very helpful and it is a time saver
17	My understanding of the contents	It's a good learning experience
18	Knowing the standard and having PD on it by Neil Marshal and also speaking about it to [the HOD] about what is required for A, M, E, etc.	It was time consuming

#	How else did you make judgements? What, other than intuition and the standard influenced your decision?	We would be grateful for any other feedback that you have about your experience with using nomoremarking for NCEA internally assessed achievement standards
19	Clarity of communication including structure; discussion with other teachers occasionally	NA
20	Clarity of response	Difficult to give feedback. Resubs? Does not save time in that case
21	Best guess	Not a sensible way to make assessor judgements
22	Compared each part of the report for each script at the same time	It would have been helpful if you could have changed your mind about which response was better as a couple of times I accidentally clicked the wrong script
23	NA	NA
24	Based on requirements of standard	When comparing scripts some were shorter than others thus both scripts had to be read through so comparing one against the other was not possible as different information was located at different places down the document. If the purpose was to compare one script to the other it would have been useful to have the sections opposite each other
25	careful reading of evidence	NA
26	My knowledge of marking the standard and the professional development offered by NZQA	NA
27	How else did you make judgements? What, other than intuition and the standard influenced your decision?	We would be grateful for any other feedback that you have about your experience with using nomoremarking for NCEA internally assessed achievement standards

References

- Alomran, M., & Chia, D. (2018). Automated scoring system for multiple choice test with quick feedback. *International Journal of Information and Education Technology*, 8, 538–545. <https://doi.org/10.18178/ijiet.2018.8.8.1096>.
- Assessment Research Group. (2009). *Assessment in schools: Fit for purpose? A commentary by the teaching and learning research programme*. London: Economic and Social Research Council.
- Baird, J.-A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles, Policy & Practice*, 24, 317–350. <https://doi.org/10.1080/0969594X.2017.1319337>.
- Berkowitz, B. W., Fitch, R., & Kopriva, R. (2000). *The Use of Tests as part of high-stakes decision-making for students: A resource guide for educators and policy-makers*. Washington, DC: Office for Civil Rights (ED).
- Bisson, M., -J., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education*, 2, 141–164. <https://doi.org/10.1007/s40753-016-0024-3>.
- Black, P., Burkhardt, H., Daro, P., Jones, I., Lappan, G., Pead, D., & Stephens, M. (2012). High-stakes examinations to support policy. *Educational Designer*, 2(5), 1–31. <https://www.educationaldesigner.org/ed/volume2/issue5/article16/>

- 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.
- Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, 37, 1–19. <https://doi.org/10.1007/BF03216919>.
- Hipkins, R., Johnston, M., & Sheehan, M. (2016). *NCEA in context*. Wellington, New Zealand: NZCER Press. <https://www.nzcer.org.nz/nzcerpress/ncea-context>.
- Hunter, J., & Jones, I. (2018). Free-response tasks in primary mathematics: a window on students' thinking. In *Proceedings of the 41st annual conference of the Mathematics Education Research Group of Australasia* (Vol. 41, pp. 400–407). Auckland, New Zealand: MERGA.
- Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, 39(10), 1774–1787. <https://doi.org/10.1080/03075079.2013.821974>.
- Jones, I., Bisson, M., Gilmore, C., & Inglis, M. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal*, 45, 662–680. <https://doi.org/10.1002/berj.3519>.
- Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics*, 89, 337–355. <https://doi.org/10.1007/s10649-015-9607->.
- 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., & Karadeniz, I. (2016). An alternative approach to assessing achievement. In C. Csikos, A. Rausch, & J. Szitanyi (Eds.), *The 40th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 51–58). Szeged, Hungary: IGPME.
- Jones, I., & Sirl, D. (2017). Peer assessment of mathematical understanding using comparative judgement. *Nordic Studies in Mathematics Education*, 22, 147–164.
- Jones, I., Swan, M., & Pollitt, A. (2014). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, 13(1), 151–177. <https://doi.org/10.1007/s10763-013-9497-6>.
- Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation*, 47, 93–101. <https://doi.org/10.1016/j.stueduc.2015.09.004>.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73. <https://doi.org/10.1111/jedm.12000>.
- Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability*. Manchester: AQA & the National Assessment Agency.
- Murphy, R. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology*, 52, 58–63. <https://doi.org/10.1111/j.2044-8279.1982.tb02503.x>.
- Newton, P. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, 22, 405–420. <https://doi.org/10.1080/0141192960220403>.
- Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Cambridge: Sage Publications.
- Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19, 281–300. <https://doi.org/10.1080/0969594X.2012.665354>.
- Sangwin, C. (2013). *Computer Aided Assessment of Mathematics*. Oxford: Oxford University Press.
- Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, 29, 211–223. <https://doi.org/10.1080/08957347.2016.1171769>.
- Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87, 245.
- Suto, W. M. I., & Nadas, R. (2009). Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education*, 24, 335–377. <https://doi.org/10.1080/02671520801945925>.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273–286. <https://doi.org/10.1037/h0070288>.
- Thurstone, L. L. (1954). The measurement of values. *Psychological Review*, 61, 47–58. <https://doi.org/10.1037/h0060035>.

- van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: Examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*, 26, 59–74. <https://doi.org/10.1080/0969594X.2016.1253542>.
- Verhavert, S., Bouwer, R., Donche, V., & Maeyer, S. D. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 26, 1–22. <https://doi.org/10.1080/0969594X.2019.1602027>.
- Wiliam, D. (2001). Reliability, validity, and all that jazz. *Education 3–13: International Journal of Primary, Elementary and Early Years Education*, 29, 17–21. <https://doi.org/10.1080/03004270185200311>.
- Wiliam, D. (2010). What counts as evidence of educational achievement? The role of constructs in the pursuit of equity in assessment. *Review of Research in Education*, 34, 254–284. <https://doi.org/10.3102/0091732X09351544>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.