frontiers in Education

ORIGINAL RESEARCH
published: 11 February 2020
doi: 10.3389/feduc.2020.00006

A Two-Stage Method for Obtaining Reliable Teacher Assessments of Writing

Stephen Humphry and Sandy Heldsinger*

Graduate School of Education, University of Western Australia, Crawley, WA, Australia

In many countries, systems have employed externally imposed large-scale standardized assessments in seeking to obtain reliable and comparable assessments. To attain the same objectives while placing value on professional judgement, appropriate methods of assessment for classroom teachers need to be developed and tested. Accordingly, this paper examines the level of reliability of assessments made by classroom teachers of narrative writing using a two-stage classroom assessment method. The students involved in the study are primary students. The results show high levels of inter-rater reliability among teachers. The findings reproduce and expand on previous evidence indicating the two-stage method is a viable method of classroom teacher assessment of written performances. Implications of the results for assessment practices in education are discussed in light of background literature, with a focus on the widely expressed desire to value professional expertise.

Keywords: teacher judgement, summative assessment, formative assessment, reliability, comparative judgement

OPEN ACCESS

Edited by:
Sharon Nichols,
University of Texas at San Antonio,
United States

Reviewed by:
Raphaël Pasquini,
University of Teacher Education
Lausanne, Switzerland
Alexandra Gunn,
University of Otago, New Zealand

****Correspondence:***
Stephen Humphry
stephen.humphry@uwa.edu.au

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *13 November 2019*
Accepted: *20 January 2020*
Published: *11 February 2020*

Citation:
Humphry S and Heldsinger S (2020) A Two-Stage Method for Obtaining Reliable Teacher Assessments of Writing. Front. Educ. 5:6. doi: 10.3389/feduc.2020.00006

INTRODUCTION

Assessment is an integral component of effective teaching and a teacher's professional judgement influences all routine aspects of their work (Du Four, 2007; Hattie and Timperley, 2007; Black and Wiliam, 2010; for example., Allal, 2013). In the last 20 years, there has been considerable work internationally to support teachers in their use of assessments to improve student learning. It is now common practice for policy documents and educational resources to define and refer to the varied aspects of assessments and to provide support and resources for teachers.

In contexts where the objective is to attain comparable assessment results across teachers, classes or schools, attaining high levels of reliability is the greater challenge for school-based assessment. In many countries across the world, reliability in assessments has largely been sought through the use of standardized assessments that are developed externally to schools and generally imposed on teachers. Although many would prefer teacher judgement to be the means of obtaining summative assessments, attaining reliable judgements has been a particular challenge (Harlen, 2004; Brookhart, 2013; Johnson, 2013).

Previous research has shown that teachers obtain reliable judgements using the method of pairwise comparison (Weber, 1834; Fechner, 1860; Thurstone, 1927). In educational contexts, the method of pairwise comparisons requires judges (normally teachers) to compare pairs of performances and judge which performance, in each pair, is of a higher quality. The terms comparative judgement, comparative pairs, and paired comparison are also used to describe the process of pairwise comparisons (Tarricone and Newhouse, 2016). The method of pairwise comparisons is distinct from assessment methods in which student performances are compared to a theoretical standard or analytic marking criteria such as rubrics (Tarricone and Newhouse, 2016).

The aim of the research reported here is to collect empirical evidence about the reliability of teacher judgements of narrative writing assessment using a specific two-stage assessment method that has been used in three previous studies (Heldsinger and Humphry, 2013; Humphry et al., 2017; Humphry and Heldsinger, 2019). In the first stage, a large sample of performances are calibrated using the method of pairwise comparison. In the second stage, teachers compare students' work to the calibrated scale to score the work. The approach aims to capitalize on the high levels of reliability typically obtained using the method of pairwise comparisons, while providing teachers with a more time-effective process as well as diagnostic information about students' performances.

The article is set out as follows. First, background is provided regarding: reliability of teacher judgements; the desire expressed to value teacher judgements; and the method of pairwise comparison applied in the study. Next, the methods employed in Stages 1 and 2 are outlined. Stage 1 applies the method of pairwise comparison to calibrate narrative performances, i.e., place them on a scale. Stage 2 involves teachers assessing student performances by comparing them against exemplars with the aid of performance descriptions. Results are then presented, with an emphasis on the levels of inter-rater reliability attained in Stage 2, which is the practical stage that can be applied by classroom teachers. Lastly, the results are discussed in light of the background and practical implications are discussed.

BACKGROUND AND LITERATURE REVIEW

Reliability

The American Psychological Association published *Technical Recommendations for Psychological Tests and Diagnostic Techniques* in 1954 which were revised in 1966 as the *Standards for Educational and Psychological Testing* and then updated in 1974, 1985, and 1999 (American Psychological Association, 1954). "One requirement that has been present since the very first version, over a half century ago, is that tests should be adequately documented, the procedures by which tests were developed should be documented, and evidence regarding the validity of the tests, and specifically the reliability, must be produced" (Black and Wiliam, 2012, p. 252).

Reliability refers to the consistency of outcomes that would be observed from an assessment if the process is repeated. Reliability is necessary but not sufficient to infer validity. Indeed, in a chapter in which they explore the concept of reliability, Black and Wiliam argue that reliability is best thought of an aspect of validity. More precisely, reliability can be thought of as being associated with the random component of construct-irrelevant variance, in that the lower this component is relative to construct relevant variance, the higher the reliability. The same authors explain that the aim of any assessment is to ensure that variations in students' scores are caused by differences that are relevant to the construct of interest, rather than irrelevant factors such as who assessed the student, the selection of assessment items, and factors that may have impeded or enhanced student performance on the day.

Discussion of reliability in the context of teachers' assessments or judgements of their students' performances tends to focus more on examiner or scorer error rather than content or sampling error, or factors that impeded or enhanced student performance. With this focus, reliability is often referred to as inter-rater reliability and is concerned with the generalizability of scores across markers or scorers. Any differences that arise in scores that are not a function of student ability, but from differences in examiners, are a source of measurement error that negatively impacts the reliability of the assessment. It is less challenging to obtain reliable teacher judgements in the context of short or closed response assessment items as there is less opportunity for examiner or scorer error. The context of extended performance assessments such as essays poses a greater challenge.

Policy documents often refer to terms such as *consistency* or *dependability* and moderation either as a proxy for, or instead of, reliability. A review of evaluation and assessment by the Organization for Economic Co-operation and Development states that "moderation refers to quality assurance measures that seek to increase the consistency of marking, for example through teachers reviewing or cross-marking each other's assessments within a school or across schools or working together in groups to discuss assessment criteria and student performance levels" (Organisation for Economic Co-operation and Development, 2013, p. 197). Policy documents from Australian education departments typically recommend that teachers moderate to improve the dependability and consistency of their judgements.

Although there is understandably some reticence to discuss reliability in technical terms when drafting policy documents for a professional rather than academic or technical audience, recommending that teachers moderate to improve the dependability and consistency of their judgements has the same intention as the technical recommendations set out in the Standards for Educational and Psychological Testing (American Psychological Association, 1999). The shared intention of the policy documents and the standards is to ensure, as far as possible, that differences in scores represent differences in the construct being measured, rather than differences caused by irrelevant factors such as who assessed the student. The important implication is that where differences in scores more accurately represent the differences in the construct being assessed, more trust can be given to scores when drawing inferences about students' ability and making decisions about follow-up actions.

Reliability of Teacher Judgement

In a 2013 article, Johnson observed that although various countries have relied to some degree on teacher judgement in high-stakes contexts, evidence about the reliability of those judgements is limited and often ambiguous. In a systematic review of the evidence for the reliability and validity of teacher judgement, Harlen (2005) identified 12 studies that had examined reliability of teacher assessment. The studies examined rescoring or remarking or moderation of the data from the assessment process, or explored the influence of school or student variables. Of the 12, only 3 studies involved experimental control or manipulation in the research. The remainder examined teacher judgement in naturally

occurring contexts or looked at the relationship between one variable and another. The review reported evidence of low reliability and bias in teacher judgements, and recommended, amongst other things, that "there needs to be research into the effectiveness of different approaches to improving the dependability of teachers' summative assessment including moderation procedures" (Harlen, 2005, p. 268).

Brookhart (2013) reviewed 100 years of study of teacher judgements in the USA and found that the research generally fell into two categories: examinations of teachers' summative grading practices and examinations of how teacher judgements of student achievement accords with large-scale summative assessment which were mostly standardized tests. The quality of teacher judgements was found to be variable in both contexts.

In Australia, the introduction of a national testing program and the concomitant reporting of school outcomes to meet public accountability demands led to an increased interest in the extent to which teacher judgements are reliable and consistent (Connolly et al., 2012). Connolly et al. (2012) analyzed interview data to identify teachers' perspectives on standards and moderation as a means of achieving consistency of teacher judgement and the authors noted that "a critical review of the research pertaining to teacher judgment reveals that teachers draw on multiple sources of knowledge and evidence when making judgements and that the use of standards, and criteria alone will not result in consistency of teacher judgements" (p. 596).

Placing Value on Teacher Judgements

In 2004, Wynne Harlen wrote that there appeared to be some willingness of policy makers in England to consider making greater use of teachers' judgement for summative assessment. Teachers' assessment was seen as a central aspect of the National Curriculum, when introduced in England and Wales in 1987. Yet a review in 1988 found that externally imposed testing had taken a stranglehold on the curriculum and teaching, and recommended parity of esteem between teachers' assessment and national test results (Harlen, 2004).

Almost a decade later, the issue had not been resolved. Johnson (2013) observed that "There is a widely felt desire within the educational community at large to value teachers' professional expertise, and, in the UK, an associated belief in the potential of teacher summative assessment to right the perceived wrongs of a controversial and unpopular testing regime that permeates the length of schooling" (p. 101). In 2015, The Commission on Assessment Without Levels, established by the UK government Department for Education, found that too great a reliance was being placed by government on external tests, particularly for school accountability purposes (Standards and Testing Agency, 2015).

Australia seems no different in that there has been an intention to make greater use of teachers' judgements, but there is little indication that the objective has been realized in the intended manner. In 2008 the Australian Ministers for Education signed the Melbourne Declaration on Educational Goals for Young Australians (MCEETYA, 2008). This declaration frames the role and purpose of education in Australia, the ministers for education committed to assessment of student progress that is rigorous and comprehensive and which draws on a combination of the professional judgement of teachers and testing, including national testing.

The ministers' commitment to develop national testing in Australia was swiftly and in 2008 the government commenced the externally implemented National Assessment Program, Literacy and Numeracy (NAPLAN). Australia has stumbled in much the same way as England, and although it's been nearly a decade since the ministers' agreement, Australia is yet to develop a rigorous and comprehensive system that draws on teacher judgement to assess student progress.

There is little indication in peer-review literature published in English that genuine gains have been made in establishing parity of esteem between teachers' assessments and externally imposed standardized tests. Given that even governments appear to want to make greater use of teacher summative judgements, it is important to understand why this has not been achieved. Possible reasons for this are foreshadowed in the two reviews of the research evidence for the reliability and validity of teachers' assessment referred to earlier: the UK review conducted by Harlen (2004) and the 100 years of study of teacher judgement in the USA conducted by Brookhart (2013).

The UK review concluded that "the findings … by no means constitute a ringing endorsement of teachers' assessment; there was evidence of low reliability and bias in teachers' judgements made in certain circumstances" (Harlen, 2005, p. 245). The USA review concluded that for "important public accountability functions, standardized tests are currently trusted over teacher judgements" (Brookhart, 2013, p. 84). The author added that "until the quality of teacher judgements in summative assessment is addressed in both research and practice, students and schools in the USA will probably continue to be addressed in a two-level manner. Teachers' grades will have immediate, localized effects on students, but for the more politically charged judgements of school accountability, standardized tests will garner more trust" (Brookhart, 2013, p. 86).

Method of Pairwise Comparisons

Background to the use of the method of pairwise comparisons in education and other fields is provided by Bramley et al. (1998), Bond and Fox (2001), Heldsinger and Humphry (2010), and Pollitt (2012). The method of pairwise comparisons is also referred to as comparative judgement; the former term emphasizes the comparison of pairs of performances. In educational contexts, the method of pairwise comparisons requires judges (normally teachers) to compare pairs of performances and judge which performance is of a higher quality. The data resulting from the judges' comparisons is then analyzed using the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 1959) to scale the performances.

The design and principle for the scale construction, in the method of pairwise comparison, is based on the work of Thurstone (1927, 1959). The scale locations are inferred from the proportions of judgements in favor of each performance when compared with others. Performances are placed on the scale from weakest to strongest. If all performances were compared with

each other, the strongest performance would be the one judged better than the other performances the greatest number of times. However, in practice, scaling techniques can be used and it is unnecessary for each performance to be compared with every other performance.

Pairwise comparisons are too time-consuming in their standard form to be viable as a general method for teacher assessment (Bramley et al., 1998). In addition, in standard applications, the method of pairwise comparison does not produce readily available diagnostic information. Pollitt (2012) has sought to attain time savings by using Adaptive comparisons. In Adaptive comparisons, the locations or scores for performances are re-estimated after successive rounds of pairwise comparisons so that in the final round of comparisons, each script is compared only to another whose current estimated score is similar. The approach attempts to increase the amount of statistical information contained in each judgement and at the same time reduce the overall number of comparisons required. Direct ranking methods have also been used, but these are challenging with even a moderate number of complex stimuli.

Heldsinger and Humphry (2013) pursued an alternative two-staged method of assessment that capitalizes on the reliability afforded by the method of pairwise comparison. The two-staged method is designed to be time-effective, accessible to classroom teachers, and informative. In Stage 1, a large number of performances are calibrated by asking teachers to compare performances and select the performance that is of a higher quality, and then analyzing their judgements using the Bradley-Terry-Luce model (Bradley and Terry, 1952; Luce, 1959). In the study reported here, 160 performances were used in Stage 1. Once all the performances have been calibrated, a qualitative analysis of the calibrated performances is used to derive empirically based descriptions of the features of development evident in the performances, and a subset of performances are selected to act as exemplars. In Stage 2, teachers assess students by comparing the students' performances to the calibrated exemplars and performance descriptors.

An earlier study examining teachers' assessments of recount writing in the early years demonstrated that a high level of inter-rater reliability can be obtained when teachers compare students' work to calibrated exemplars. The average inter-rater correlations by judge ranged from 0.897 to 0.984 (Heldsinger and Humphry, 2013). In this early work, performance descriptors were not provided as a complement to the exemplars. The reliability of judgements using this method, and including performance descriptors, were subsequently investigated for oral language in early childhood (Humphry et al., 2017) and for essay writing in primary students (Humphry and Heldsinger, 2019). The present study extends the previous research to investigate the reliability of the methodology in the context of narrative writing obtained from students aged 5–12.

METHOD

Overview

In Stage 1, a large sample of performances is compared in pairs by a group of teachers to create a scale. A subset of exemplars is selected to act as anchor points on the scale that is constructed. Performance descriptors are then developed to complement the exemplars. In Stage 2, classroom teachers assess a common set of narrative performances by comparing them to the set of calibrated exemplars and accompanying performance descriptors. Teachers decide whether a performance is of a similar level to an exemplar, in order to place it on the scale at the level of the exemplar. The performance descriptors assist the teachers in understanding the features of development evident in the exemplars, but the comparison to the exemplars is the critical component of the assessment process.

The concept of using exemplars to support reliable judgements has been explored for some time. For example, a 1965 discussion pamphlet prepared by members of the London Association for the Teaching of English described an assessment process in which 28 imaginative compositions by 15-year-old students were arranged in order of merit from the most inadequate to the best, and each was accompanied by a commentary. The distinctive aspect of the assessment method described in this study is the systematic calibration of student performances using pairwise comparisons to create a scale and ordering. The application of pairwise comparisons also affords technical and practical advantages, such as enabling tests of internal consistency.

Stage 1: Design and Development of the Narrative Writing Assessment

The first stage focused on the calibration of the exemplars to be used in the second stage of this study. Stage 1 replicated previous studies by Heldsinger and Humphry (2010, 2013) in the context of narrative writing of students in primary schools.

Participants

The first stage was conducted in five government Western Australian primary schools in the both metropolitan (four) and regional (one) areas. Due to the limited number of schools in this study, it was not feasible to employ a full stratified or other sampling design. Nevertheless, to obtain some heterogeneity, the schools were selected to reflect a mix of socioeconomic contexts as indicated by their values on the Index of Community Socio-educational Advantage (ICSEA; $M = 999$, $SD = 93$). This index (ICSEA) is used in the Australian National Assessment Programme—Literacy and Numeracy (NAPLAN). The median for Australian schools is 1,000 and the standard deviation is 100 (Australian Curriculum Assessment and Reporting Authority, 2017).

Participating teachers collected the narrative performances in term 1 from 160 children (gender was not identified) in Pre-primary to Year 7. Children were aged from 5 to 12 years of age.

Task Administration

The assessment task was devised by the researchers in consultation with expert teachers and it consisted of narrative writing prompts and administration guidelines. Care was taken to ensure that the prompts would elicit narrative writing, as opposed to recount or persuasive writing for example.

Teachers were provided with the following prompts: (1) What a discovery! This would change things for sure.; (2) A loud noise woke you. As you opened your door. . . .; (3) You will never believe this but I have to tell you anyway; (4) It was gone! Peeking through the window, she now knew what she had to do.; (5) Nothing is even as it seems!; and (6) I woke up with a start. Something was shining through my window. They could choose from these prompts or provide one of their own as long as the task elicited narrative writing.

The task was administered individually by classroom teachers. The administration was not strictly standardized, however, teachers were given administration instructions as a guide, which they could follow or adapt where appropriate. This included a guide for time allocation, narrative topics and instructions to read to the students.

Pairwise Comparisons

A total of 23 teachers, from 12 metropolitan and regional schools, participated as judges in this study. Some of the teachers were involved in collecting the narrative samples. All were experienced classroom teachers. Each received 30 min of training to make holistic judgements about students' written narrative skills and to use the assessment, and reporting software to make and record judgements. In making this judgement they were asked to consider both authorial choices and conventions. Authorial choices include such aspects as subject matter, language choices, character and setting, and the reader-writer relationship. Conventions includes aspects where the writer is expected to largely follow rules including spelling, punctuation, correct formation of sentences, and clarity of referencing.

Judges were given online access to specific pairs of narrative performances to be compared. The pairs were generated randomly from the list of all pairs of performances. Judges worked individually and compared pairs of performances based on holistic judgements as to which performance displayed more advanced writing skill. Each judge made between 3 and 200 comparative judgements. A total of 3,532 pairwise comparisons were made as a basis for forming the scale. For further detail on how the method yields exemplars located on the scale used in Stage 2, see Heldsinger and Humphry (2010, 2013).

The Selection of Calibrated Exemplars and Development of Descriptors

First, from the original 160 performances, a subset of 17 performances was selected as exemplars. Care was taken to select exemplars that most clearly and typically captured developmental features at given points on the scale.

Second, a descriptive qualitative analysis based on all 160 narrative performances on the scale was undertaken to describe the features of narrative writing development. Qualitative analysis examined both aspects of writing (authorial choices and the conventions). This led to the drafting of performance descriptors and teaching points based on the empirical data, given the ordering of performances established by pairwise comparisons. In combination, the exemplars and the performance descriptors characterized the development of narrative writing in a manner intended to enable the assessment of separate performances in Stage 2.

Stage 2: Assessment of Student Performances Against Calibrated Exemplars

In Stage 2, two groups participated for the purpose of investigating the reliability of the two-staged assessment process. The first group comprised of practicing classroom teachers invited to participate. The second group were self-selected practicing classroom teachers and education consultants who participated as part of a certification process.

From the 160 performances in the first stage, 25 performances were selected as common performances to assess in stage 2. The 25 performances had a mean and standard deviation, after transformation, of 340.7 and 69.0, respectively. In Stage 2, the teachers assessed the 25 common narrative performances by comparing the performances with the calibrated exemplars. As judges they had to decide whether a student's performance was qualitatively similar to an exemplar, or whether it fell between two exemplars, and then score the performance accordingly. **Table 1** provides an overview of the design of Stage 2.

Exemplars and Performance Descriptors

The 17 exemplars and descriptors were displayed adjacent to a vertical scale in a customized web application, for which the assessment display is shown in **Figure 1**. In this display, performances to be assessed appear on the right-hand side and descriptors appear on the left-hand side. Thumbnails of the 17 calibrated exemplars appear adjacent to the scale in the center.

The judges were asked to make an on-balance judgement based on their analysis of the strengths and weaknesses of the performance, and to determine which exemplar the performance was closest to or which two exemplars it fell between. The judges had the option to make one of three comparisons: the sample was exactly the same level (in terms of writing ability) as the exemplar, or it was slightly better than the exemplar, or it was slightly weaker than the next exemplar, or it fell halfway between both exemplars. They scored the performance accordingly.

The judges were provided with a guide to help make their judgements. This guide contained all the calibrated exemplars, the performance descriptors, and a close qualitative analysis of each exemplar. It was designed to help participants familiarize themselves with the exemplars and understand the particular features of each.

RESULTS

Stage 1: Analysis of Pairwise Data

A total of 23 primary classroom teachers compared a total of 160 performances. A total of 3,532 comparisons was made. The Person Separation Index was 0.963, indicating a high level of internal consistency (Andrich, 1988; Heldsinger and Humphry, 2010). The Person Separation Index is an index of internal

TABLE 1 | Overview of the design of Stage 2.

Component	*N*	Background information
Student writing samples	25	25 written narrative performances were selected from the original 160 performances used in the pairwise comparisons, excluding any performances that had been selected as calibrated exemplars. The scale locations had a mean of 0 and standard deviation of 4.95. Performances were stratified into the lowest third, middle third and highest third in terms of scale ranges and a selection of 8, 9, and 8 performances, respectively were made from these strata.
		The same 25 performances were used by the group 1 and group 2 judges.
Group 1 judges	12	Twelve judges, none of whom had participated in Stage 1. All participants were practicing classroom teachers.
Group 2 judges	37	Thirty-seven judges, none of whom had participated in Stage 1 or group 1.
		Thirty-four judges were practicing classroom teachers and three judges were expert markers who are responsible for developing training materials.

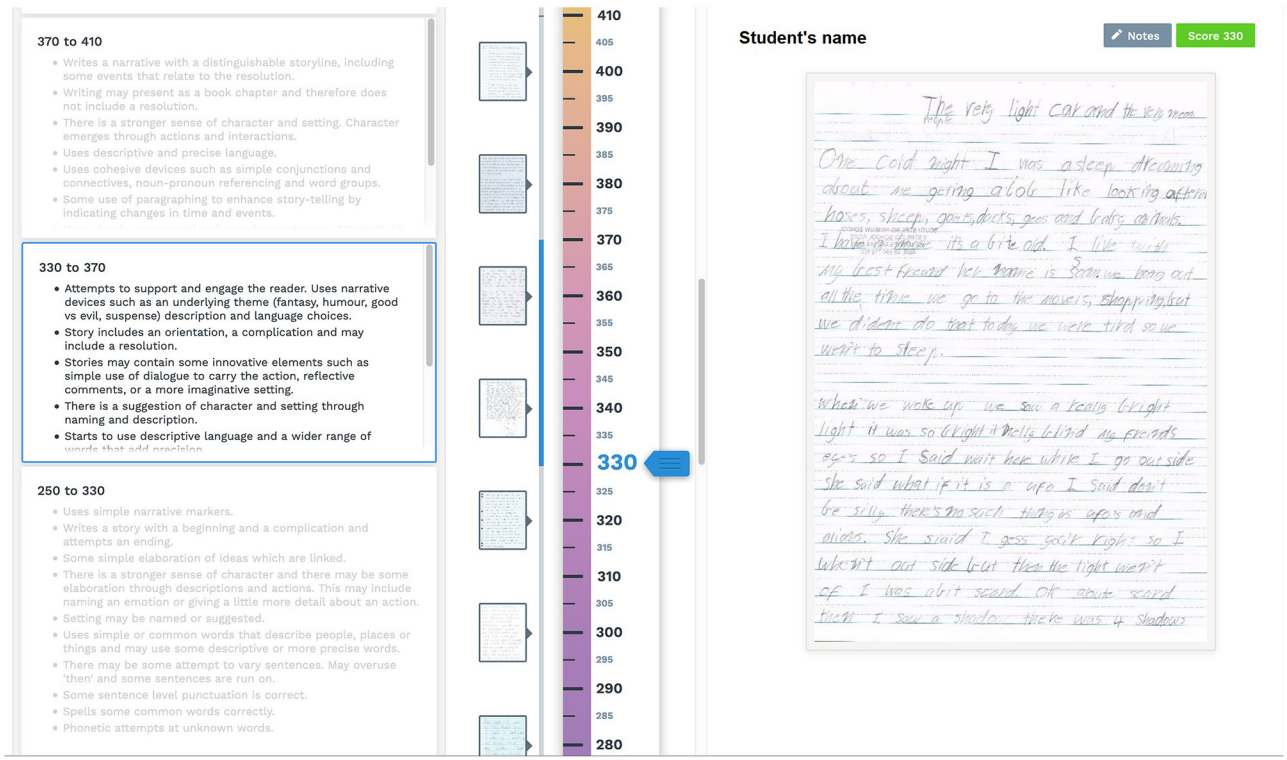

FIGURE 1 | Stage 2 assessment display. Note that a linear transformation was applied to the scale obtained from the analysis of pairwise data so that the exemplars had a range of ∼110 to 300. This is so the range is not interpreted as a percentage, and also to avoid negative scale locations and decimal places. This kind of transformation is commonly used in large-scale testing programs.

reliability used in Item Response Theory that is based on and analogous to Cronbach's alpha.

After applying the BTL scaling model, the original mean and SD for the scale locations were 0 and 3.418, respectively. A linear transformation was made such that the mean and SD were 330 and 80, respectively to make the range more readily interpretable for classroom teachers by avoiding negative numbers and decimal places. This transformation constitutes an arbitrary change of the unit and origin of the original interval scale obtained by application of the model. From the 160 performances, a subset of 17 performances at equal intervals of 20 from 160 to 500 were selected as exemplars.

Stage 2

The following results are presented for groups 1 and 2 separately: (a) summary of bivariate inter-rater correlations; (b) rater-average correlations; and (c) rater harshness. Rater-average correlations provide a more straightforward summary of reliability levels, but inter-rater reliability based on pairs of raters are used fairly commonly in applied settings and provide another point of reference for interpretation of results.

Inter-rater Reliability

Inter-rater reliability is here defined as the correlation between a pair of raters. For group 1, the average inter-rater correlation

TABLE 2 | Group 1 rater reliability and harshness indicators.

Judge	Rater-average correlation	Harshness
J1	0.847	−31.5
J2	0.918	−4.1
J3	0.888	15.3
J4	0.835	−9.0
J5	0.969	8.0
J6	0.973	−2.5
J7	0.813	−17.8
J8	0.837	22.7
J9	0.938	−15.1
J10	0.957	8.0
J11	0.932	9.1
J12	0.921	16.9

TABLE 3 | Group 2 rater reliability and harshness indicators.

Judge	Rater-average correlation	Harshness
J1	0.973	10.6
J2	0.951	9.0
J3	0.913	18.8
J4	0.939	−2.9
J5	0.939	−4.6
J6	0.975	11.7
J7	0.878	6.1
J8	0.813	−20.1
J9	0.919	6.3
J10	0.936	−7.1
J11	0.969	0.3
J12	0.943	−7.1
J13	0.912	15.9
J14	0.893	−1.1
J15	0.977	11.0
J16	0.948	4.3
J17	0.948	−2.2
J18	0.967	13.0
J19	0.948	5.4
J20	0.865	−1.1
J21	0.907	−17.6
J22	0.938	19.5
J23	0.892	10.3
J24	0.894	−29.4
J25	0.893	−22.5
J26	0.935	7.2
J27	0.944	−13.1
J28	0.920	19.0
J29	0.940	−4.4
J30	0.933	−31.0
J31	0.935	−3.1
J32	0.913	−9.1
J33	0.848	−10.9
J34	0.971	0.9
J35	0.893	−19.4
J36	0.968	9.7
J37	0.967	27.5

between each of the 65 unique pairs of raters was 0.835 (range 0.614 to 0.972). For group 2, the average inter-rater reliability was 0.863 (range 0.602 to 0.997). These indicate a generally high level of inter-rater reliability with some variability.

Rater-Average Correlation

Tables 2, **3** show the correlation between each rater's scores and the average scores of all other raters (rater-average correlations) for groups 1 and 2, respectively. The rater's score for a performance was compared with the mean score from all other judges by excluding the rater's own score from the mean. Then the correlation between these two scores, for all 25 common performances, was used to compute the rater-average correlation for each rater. In group 1, the mean rater-average correlation was 0.902 (range from 0.813 to 0.973), as shown in **Table 2**, which again indicates a high level of inter-rater reliability. In group 2, the mean rater-average correlation was 0.927 (range from 0.813 to 0.977). These correlations again indicate generally high levels of reliability with variability across raters as shown in **Table 3**.

Rater Harshness

An indicator of rater harshness is obtained by taking the average of a given rater's scores for the common performances and subtracting the overall average for all the other raters. A positive score means the rater awarded higher scores on average to the performances (lenient), a lower score means the rater awarded lower scores on average (harsh). The harshness indicator appears in **Table 2** for group 1 and in **Table 3** for group 2. There were some relatively lenient and harsh raters. The standard deviation of rater harshness levels was 16 in group 1 and 14 in group 2 (the standard deviation of the average scores of 64 in both groups). Variation in rater harshness was relatively modest with the largest outliers being harsh raters (rater 1 in group 1, and raters 24 and 30 in group 2).

DISCUSSION

The study aimed to examine the reliability of teachers' assessment based on a two-stage method of assessment designed to capitalize on the reliability of pairwise comparisons with increased time-effectiveness and diagnostic information. In doing so, the aim of this research is to test and extend the results of earlier research that found that the two-staged method provided reliable teacher assessments in the context of early childhood recount writing. The study accords with Brookhart (2013) 's recommendation for further research into the effectiveness of different approaches to improving the dependability of teachers' summative assessment.

In Stage 1, the method of pairwise comparisons was used to calibrate a large sample of 160 narrative performances. The high reliability obtained is similar to findings from earlier studies using the paired comparison process cited earlier in the paper. The findings from Stage 2 showed that high inter-rater reliability

was achieved when teachers used calibrated exemplars and performance descriptors to assess performances, establishing the viability of the second stage for classroom teachers to achieve reliable assessments.

Although the teachers in this study were not assessing their own students' work, they were assessing student performances that are typical of those collected in primary classroom contexts. In addition, the assessment methodology required no specific training, other than time needed for teachers to become familiar with the exemplars and accompanying descriptors in the custom software. The teachers and consultants in group 2 who participated as part of a certification process achieved slightly higher levels of reliability than the teachers recruited to participate in group 1. This is not surprising given participants in group 2 were self-selecting whereas, teachers in group 1 were invited to participate. Group 2 results indicate attainable levels of reliability and group 1 results provide verification that the levels of reliability are relatively similar in a group invited to participate and therefore, not self-selected.

Reliability was examined using inter-rater correlations for all pairs of raters as well as rater-average correlations. If raters tend to assess performances in a fairly consistent manner, the averages provide an estimate of the scale location for a performance that has less measurement error than that associated with an individual rater's assessment. Consequently, rater-average correlations tend to be higher than rater-rater correlations. Rater-average correlations are a more straightforward summary, however these are only available when all raters mark a common set of scripts. Rater-rater correlations are fairly common in applied settings and they are reported to facilitate interpretation of the levels of reliability obtained in this study for readers accustomed to this indicator.

Comparing the findings with those of previous studies, similar levels of inter-rater reliability were obtained in Stage 2 as those found in the study by Heldsinger and Humphry (2013) focusing on the assessment of early childhood recount writing, using a paper version of the two-stage method of assessment. However, the findings of this study were obtained from older students who wrote longer texts and, in the context of narrative writing, which could be considered more demanding than recount writing. The findings of this study are also very similar to those reported by Humphry and Heldsinger (2019) who used the same two-stage method for assessing persuasive essay writing.

The variations in rater harshness observed in Stage 2 of the research are generally modest relative to the variation in the student performances (on the same scale). However, in both groups there were outliers who were notably harsh, giving average scores ∼30 points lower than the overall average of all other raters. This could have implications for feedback to students and for evaluation of class-level results in which a harsh rater assesses all students in a classroom. The variations in rater harshness might be reduced with feedback on practice performance. Further research would be required to ascertain whether this is effective and feasible.

The two-staged method was developed as a more time-effective process than pairwise comparisons and one which allows for the provision of diagnostic information about students' work. The work undertaken in Stage 1 to calibrate a scale need only be undertaken once by a relatively small number of teachers and once calibrated the scale in available for all teachers to use.

In addition, the assessment process used in Stage 2 is comparable to the time taken for assessment of similar performances in large-scale programs for which one of the authors has overseen the training. Across the groups, it took ∼7.4 min per assessment for teachers familiar with the exemplars and descriptors to assess a performance, with an average of 24.2 assessments made in an allotted 3 h. In large-scale testing, training is required. In this context, familiarity with exemplars is required but minimal training is entailed. In addition, the more teachers assess using the calibrated exemplars and descriptors, the more familiar they become with exemplars and the more quickly it may be expected that they will make judgements.

A motivation for using teacher assessment instead of or to complement other tests is to recognize, develop and value the professionalism of teachers, which is consistent with a desire in education noted by Johnson (2013). The results of this study also provide empirical evidence that the approach is promising for obtaining dependable teachers' summative assessment in line with calls for research into the reliability of teacher assessments by Harlen (2005), Brookhart (2013), and Johnson (2013). In this context the teachers were not assessing their own students' work, however, the approach readily translates to such contexts provided there are not biases involved with assessing students known to the teachers.

It has been shown that a negative impact of the higher profile given to test-based results in England's national curriculum assessment system was not only a loss of assessment skill on the part of teachers, but also a loss of confidence in their ability to make sound assessments of their students (Black et al., 2010, 2011). The findings from this study are promising in that they provide one potential means for restoring such confidence. The levels of reliability indicate that teacher judgements can be trusted and valued when the two-staged assessment method is adopted for narrative writing.

Given the similarity of the assessment task, and the overlap in age range of participating students, a direct comparison with inter-rater reliability in NAPLAN would be instructive. However, to our knowledge, levels of inter-rater reliability are not reported for NAPLAN and therefore, it was not possible to make this comparison. However, the rater-average correlations are very similar to unpublished results obtained by the authors during marker training, using a similar design, in a precursor to NAPLAN, the Western Australian Literacy and Numeracy Assessment program. The NAPLAN marking guide is based largely on the Western Australian guide (Humphry and Heldsinger, 2014).

The research is limited to the context of primary schools. Further research is required to ascertain levels of reliability of the two-staged method applied in contexts where performances are longer, such as secondary schools and tertiary contexts. The maximum length of the performances in this study was four pages. Further research is also required to examine the reliability of the two-stage assessment method in different learning areas such as History or Visual Arts. The research is also limited to the

context of summative assessment. Further research is required to better understand the extent to which obtaining reliable teacher assessment of writing using the two-staged method of assessment can inform and possibly enhance teachers' formative assessment processes.

CONCLUSION

In summary and conclusion, Stage 1 of the study involved conducting pairwise comparisons of 160 narrative performances for which there was a high level of internal consistency and good fit to the model of analysis. This is similar to previous findings related to the application of pairwise comparisons for essays.

In Stage 2, a number of teachers used calibrated exemplars and empirically derived performance descriptors to assess a common set of 25 performances. A high level of inter-rater reliability was obtained. There was variation in the level of reliability and also some variation in rater harshness using the method. The method requires minimal training to obtain good reliability.

The findings indicate that the two-staged method of assessment holds promise for enabling primary teachers to make reliable judgements of narrative writing. An advantage of the two-stage method over pairwise comparisons alone is that once a scale has been constructed, the average time to assess a performance is reasonably modest. Also, unlike in externally imposed testing programs, classroom teachers assess their own students and can provide formative feedback based on their assessments, with reference to performance descriptors and performance exemplars as needed.

This study responds to the need noted by Harlen (2005) for investigation of the effectiveness of different approaches to improving the dependability of teachers' judgements. The two-stage method provides a clear and explicit basis for moderation of teacher judgements. The results suggest that the method of assessment employed by classroom teachers is important to the quality of the results obtained, and that care needs to be taken in designing an appropriate method. The results indicate that the two-stage method is an effective method by which to meet the widespread desire in education to value professional expertise, observed by Johnson (2013), by enabling teachers to make dependable and comparable judgements of performance.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the University of Western Australia. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Both authors contributed to the conceptualization and drafting of the manuscript. SHu focused specifically on technical aspects of research design and data analysis. SHe focused specifically on co-ordination of the research project.

FUNDING

This work was supported by an Australian Research Council Linkage grant (LP140100567) with the Australian Curriculum and Standards Authority (ACARA), the School Curriculum and Standards Authority (SCSA), and the Board of Studies (NSW) as Partner Organizations.

ACKNOWLEDGMENTS

Agreement by the School Curriculum and Standards Authority, Western Australia was acknowledged for the use of the empirical data featured within the paper.

REFERENCES

Allal, L. (2013). Teachers' professional judgement in assessment: a cognitive act and a socially situated practice. *Assess. Educ. Princ. Policy Pract.* 20, 20–34. doi: 10.1080/0969594X.2012.736364

American Psychological Association (1954). *Technical Recommendations for Psychological Tests and Diagnostic Techniques. Developed jointly by American Psychological Association, American Educational Research Association, & National Council on Measurement in Education*. Washington, DC: American Psychological Association.

American Psychological Association (1999). *Standards for Educational and Psychological Testing, 3rd Edn. Developed jointly by American Psychological Association, American Educational Research Association, & National Council on Measurement in Education*. Washington, DC: American Psychological Association.

Andrich, D. (1988). *Rasch Models for Measurement*. Beverly Hills, CA: Sage Publications. doi: 10.4135/9781412985598

Australian Curriculum Assessment and Reporting Authority (2017). *Guide to understanding 2013 Index of Community Socio-educational Advantage (ICSEA) values*. Retrieved from Australian Curriculum Assessment and Reporting Authority website: https://acaraweb.blob.core.windows.net/resources/Guide_to_understanding_2013_ICSEA_values.pdf (accessed January 30, 2020).

Black, P., Harrison, C., Hodgen, J., Marshall, B., and Serret, N. (2010). Validity in teachers' summative assessments. *Assess. Educ. Princ. Policy Pract.* 17, 217–234. doi: 10.1080/09695941003696016

Black, P., Harrison, C., Hodgen, J., Marshall, B., and Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assess. Educ. Princ. Policy Pract.* 18, 451–469. doi: 10.1080/0969594X.2011.557020

Black, P., and Wiliam, D. (2010). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan* 92, 81–90. doi: 10.1177/003172171009200119

Black, P., and Wiliam, D. (2012). "The reliability of assessments," in *Assessment and Learning, 2nd Edn*, ed J. Gardner (London: Sage Publications Ltd.), 243–263.

Bond, T. G., and Fox, C. M. (eds.). (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc. doi: 10.4324/9781410600127

Bradley, R. A., and Terry, M. E. (1952). Rank analysis of incomplete block designs, I. The method of paired comparisons. *Biometrika* 39, 324–345. doi: 10.1093/biomet/39.3-4.324

Bramley, T., Bell, J. F., and Pollitt, A. (1998). Assessing changes in standards over time using Thurstone's paired comparisons. *Educ. Res. Perspect.* 25, 1–24.

Brookhart, S. M. (2013). The use of teacher judgement for summative assessment in the USA. *Assess. Educ. Princ. Policy Pract.* 20, 69–90. doi: 10.1080/0969594X.2012.703170

Connolly, S., Klenowski, V., and Wyatt-Smith, C. (2012). Moderation and consistency of teacher judgement: teachers' views. *Br. Educ. Res. J.* 38, 593–614. doi: 10.1080/01411926.2011.569006

Du Four, R. (2007). "Once upon a time: a tale of excellence in assessment," in *Ahead of the Curve. The Power of Assessment to Transform Teaching and Learning,* ed D. Reeves (Indiana: Solution Tree Press), 253–267.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel.

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

Harlen, W. (2005). Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Res. Pap. Educ.* 20, 245–270. doi: 10.1080/02671520500193744

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Heldsinger, S. A., and Humphry, S. M. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *Aust. Educ. Res.* 37, 1–19. doi: 10.1007/BF03216919

Heldsinger, S. A., and Humphry, S. M. (2013). Using calibrated exemplars in the teacher-assessment of writing: an empirical study. *Educ. Res.* 55, 219–235. doi: 10.1080/00131881.2013.825159

Humphry, S. M., and Heldsinger, S. (2019). A two-stage method for classroom assessments of essay writing. *J. Educ. Meas.* 56, 505–520. doi: 10.1111/jedm.12223

Humphry, S. M., Heldsinger, S., and Dawkins, S. (2017). A two-stage assessment method for assessing oral language in early childhood. *Aust. J. Educ.* 61, 124–140. doi: 10.1177/0004944117712777

Humphry, S. M., and Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educ. Res.* 43, 253–263. doi: 10.3102/0013189X14542154

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Res. Pap. Educ.* 28, 91–105. doi: 10.1080/02671522.2012.754229

Luce, R. D. (1959). *Individual Choice Behaviours: A Theoretical Analysis*. New York, NY: J.Wiley.

MCEETYA (2008). *Melbourne Declaration on Education Goals for Young Australians. Ministerial Council on Education, Employment, Training and Youth Affairs*, *December 30*. Retrieved from: http://apo.org.au/node/29859 (accessed January 30, 2020).

Organisation for Economic Co-operation and Development (2013). *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*. Paris: OECD Publishing. doi: 10.1787/9789264190658-en

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assess. Educ. Princ. Policy Pract.* 19, 281–300. doi: 10.1080/0969594X.2012.665354

Standards and Testing Agency (2015). *Government response: Commission on Assessment Without Levels. Standards and Testing Agency, September 17*. Retrieved from: https://www.gov.uk/government/publications/commission-on-assessment-without-levels-government-response (accessed January 30, 2020).

Tarricone, P., and Newhouse, C. P. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. *Int. J. Educ. Technol. High. Educ.* 13:16. doi: 10.1186/s41239-016-0018-x

Thurstone, L. L. (1927). A law of comparative judgment. *Psychol. Rev.* 34, 273–286. doi: 10.1037/h0070288

Thurstone, L. L. (1959). *The Measurement of Values*. Chicago, IL: The University of Chicago Press.

Weber, E. H. (1834). "De Tactu," in *De Pulsu, Resorptione, Auditu, et Tactu. Annotationes Anatomicae et Physiologicae* (Leipzig, Germany: C. F. Koehler), 44–174.

Conflict of Interest: The authors developed software to make the two-stage assessment process widely available to schools by commercializing intellectual property, assigned in part by The University of Western Australia. Licenses to the software have been purchased by government agencies and schools. The authors therefore declare that the research was conducted in the presence of a conflict of interest.