frontiers | Frontiers in Education

TYPE Methods
PUBLISHED 02 August 2023
DOI 10.3389/feduc.2023.1183695

OPEN ACCESS

EDITED BY
Janet Clinton,
The University of Melbourne, Australia

REVIEWED BY
Stefan Daniel Keller,
University of Teacher Education Zuerich, Switzerland
John Hattie,
The University of Melbourne, Australia

*CORRESPONDENCE
Carolyn Maxwell
✉ carolyn.maxwell@research.uwa.edu.au

RECEIVED 10 March 2023
ACCEPTED 20 July 2023
PUBLISHED 02 August 2023

CITATION
Humphry S, Adie L, Maxwell C and Sappl S (2023) Using pairwise comparison and ordered exemplars as a basis for a novel method of standard setting in narrative writing. *Front. Educ.* 8:1183695.
doi: 10.3389/feduc.2023.1183695

Using pairwise comparison and ordered exemplars as a basis for a novel method of standard setting in narrative writing

Stephen Humphry[1], Lenore Adie[2], Carolyn Maxwell[1]* and Sonia Sappl[1]

[1]Graduate School of Education, University of Western Australia, Perth, WA, Australia, [2]Institute for Learning Sciences & Teacher Education, Australian Catholic University, Brisbane, QLD, Australia

Standard setting of extended performances is fraught with difficulties due to the subjective nature of the scoring decision. This article introduces and reports results of a novel systematic method for setting standards for assessments that require extended performances: the Scaled Exemplar Standard Setting method (SESS). This two-stage method brings together (1) pairwise comparison of samples to scale performances and (2) standard setting involving multiple standards with tasks presented to judges in a validated order. By utilizing paired comparison methodology in the first stage, the judgment is made more manageable by chunking it into a series of pairs. This method produces scaled performances that can be checked for internal consistency of the judgments. The validated, ordered performances are used in the standard setting stage, making the task for judges as transparent and straightforward as possible. Recommendations and implications for standard setting are discussed in light of the results.

KEYWORDS

assessment, teacher judgment, standard setting, pairwise comparison, extended performance, scaled exemplars, reliability

Introduction

This research focuses on the use of a standard setting method for extended performances. Much of the literature on standard setting focuses on the application of procedures to tests comprising short-response items. The most widely used of these are the Angoff method (Angoff, 1971), the modified Angoff method, and the Bookmark method (Lewis et al., 1996). Standard setting methodologies such as these utilize Item Response Theory (IRT) and typically involve participants judging the probabilities of success on questions. However, judges do not accurately estimate absolute probabilities of success on individual questions very well (Lorge and Kruglov, 1953; Shepard, 1995; Impara and Plake, 1997, 1998; Humphry et al., 2014). In the context of standard setting methodologies, a key advantage of the method introduced in this article is that it does not involve humans judging probabilities of success on tasks.

The Angoff and Bookmark methods have also been found to be not readily applicable to extended performance (Hambleton et al., 2000). These authors identified a long-standing need for novel valid methods that are more suitable. Plake and Hambleton (2000) reported the use of a categorical standard-setting procedure applied to extended performance assessments, which highlighted the need for refinements in the outcomes of judgments when setting standards. However there has been relatively little research on standard setting for extended performances since that time. More directly relevant to this article is research by Wyatt-Smith et al. (2020) focusing on standard-setting of extended performances on a teaching performance assessment: the Graduate Teacher Performance Assessment (GTPA). The approach described in this article is similar to that used in the GTPA study in that both studies utilize pairwise comparisons as a foundational step. However, the present research uses a somewhat different procedure for standard setting and focuses on multiple standards as opposed to a single standard, extending work conducted in this field.

This article describes and reports results of the novel Scaled Exemplar Standard Setting method (SESS). The aim is to have a systematic method for setting standards for assessments that require extended performances. Ordering by paired comparisons is more manageable because the task is chunked down into a series of pairs. In addition, paired comparisons enable scaling, and the internal consistency of judgments that result in the ordering can be checked. Having methodically-ordered performances makes the task for judges as straightforward as possible by presenting extended performances transparently in a validated order.

The paper first briefly describes the materials and equipment used in each stage of the SESS method. Then the background to the project in which the SESS method was designed and trialed is outlined as well as the rationale for the study. Next, the two stages of the method are each situated within the relevant literature and discussed in terms of their application in the field of education. Details of the methods, including the design and development of the project for each of the two stages, are discussed separately, and the procedures conducted in each stage are set out. The results are presented, followed by acknowledgement of the limitations. Lastly, there is discussion of the findings, including the advantages of using this two-stage method for setting standards in extended performances in education, as well as the results from this illustrative study that have more general relevance to applications of the method.

Materials and equipment

Pairwise comparison and standard setting processes are conducted on purpose-designed software applications. The first of these enables judges to compare performances presented on screen and to select which they consider to be better according to specified criteria (Figure 1). Second, we use customized software to conduct the analysis of the data collected from the paired comparison application. We then use a web application to present multiple scaled performances and descriptions of standards with explanatory notes to participants (Figure 2). Participants were provided with instructional videos and written explanations of how to conduct paired comparisons and subsequently how to select the sample they deemed to best represent a given standard, e.g., *C* standard.

Methods

Background

The SESS method was designed within a broader study investigating the use of scaled exemplars in online moderation. The project utilized a mixed-method research design focusing on the middle years of schooling and the disciplines of English, Science, Mathematics, and Religious Education across two Australian states (Queensland and Western Australia). Over 1,400 student assessment samples were collected from over 100 teachers. From this group of assessments, 695 de-identified samples were selected for inclusion in pairwise comparison activities. Selected samples represented a quality range from highest to lowest performance (*A* to *E*), as well as coverage of school regions, school size, and socio-economic indicators. Teachers worked online to judge these samples in each discipline using pairwise comparison followed by a standard-setting process to identify samples that best represented a standard descriptor. They then wrote a commentary of how their judgment decision was made for the selected exemplars and met online with other year level and discipline teachers to discuss and refine the commentaries. The exemplars with the accompanying commentaries were trialed to ascertain their effectiveness to support teacher moderation of their judgment decisions on their own assessments.

This paper describes the processes of pairwise comparison and standard setting for English narrative writing.

Stage 1

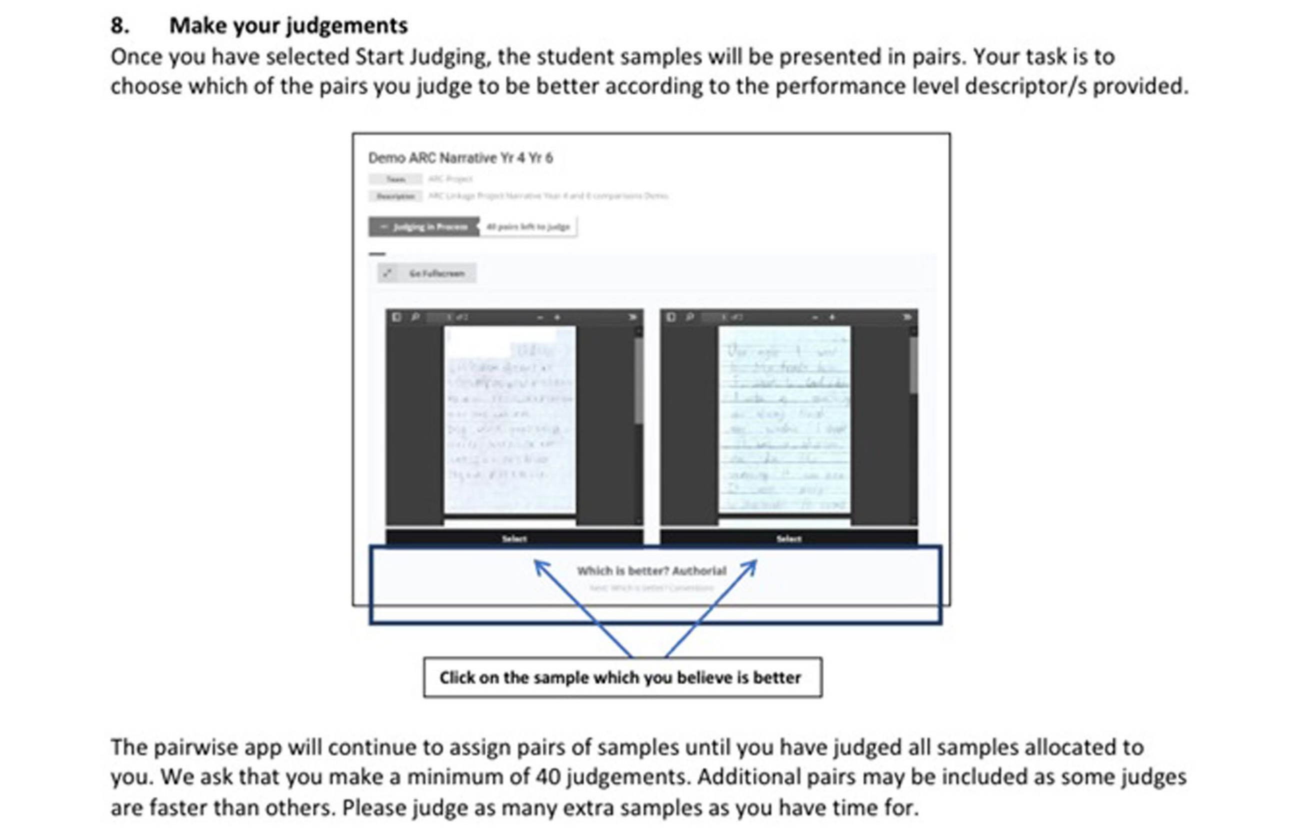
Stage 1 of this two-stage standard setting process involved teachers in a pairwise comparison activity. The method of pairwise comparison, originally developed by Thurstone (1927), involves a judgment about which of two presented samples is the better performance. Thurston showed that paired comparison of stimuli could be used to develop a scale (Humphry et al., 2017). The method has been used in a range of fields, including education (Bramley et al., 1998; Bond and Fox, 2001; Heldsinger and Humphry, 2010).

There is increasing interest in the use of pairwise comparison in education across a range of disciplines, particularly in the development of scales for writing performances (Steedle and Ferrara, 2016; van Daal et al., 2016; Humphry and Heldsinger, 2019) and mathematical problem solving (Jones et al., 2015; Jones and Inglis, 2015; Bisson et al., 2016). Others have investigated the use of pairwise comparison for peer assessment and feedback (Seery et al., 2012; Potter et al., 2017), creative performance assessments (Tarricone and Newhouse, 2016), and oral narrative performance assessments (Humphry et al., 2017).

Pairwise comparison theoretically mitigates the scope for marker harshness because performances are directly compared against each other not a separate rubric (Andrich, 1978). The method has also been shown to produce reliable results (Steedle and Ferrara, 2016). Other researchers who have adopted this approach for scaling exemplars have obtained high to very high internal consistency (Benton and Elliot, 2016; Steedle and Ferrara, 2016; Wyatt-Smith et al., 2020). However, as a method of judgment, pairwise comparisons are considered to be time-consuming (Bramley et al., 1998) with limited direct formative information on a performance (Humphry and Heldsinger, 2019). To utilize the strengths of pairwise comparison while addressing the limitations, the present study introduced a second stage involving standard setting and developed resources to support comparable judgment making.

Stage 2

The second stage shifts teachers' attention from paired comparison between samples to 'comparison' between samples and performance



8. Make your judgements

Once you have selected Start Judging, the student samples will be presented in pairs. Your task is to choose which of the pairs you judge to be better according to the performance level descriptor/s provided.

Demo ARC Narrative Yr 4 Yr 6
Team ARC Project
Description ARC Linkage Project Narrative Year 4 and 6 comparisons Demo
Judging in Process
40 pairs left to judge
Go Fullscreen
Select
Select
Which is better? Authorial
Next: Which is better? Conventions

Click on the sample which you believe is better

The pairwise app will continue to assign pairs of samples until you have judged all samples allocated to you. We ask that you make a minimum of 40 judgements. Additional pairs may be included as some judges are faster than others. Please judge as many extra samples as you have time for.

FIGURE 1
Graphic of screen and instructions for pairwise comparison in Stage 1.

level or standard descriptors. This is a process of selection in which participants choose the sample that best matches the standard descriptor.

It is well known in the judgment literature that standards written as verbal descriptors can be interpreted differently by judges (Sadler, 1987; Smith, 1989, 1995; Hudson et al., 2017; Wyatt-Smith et al., 2020).[1] Both statistical and qualitative approaches to improving and ensuring judgment consistency have been suggested, with limitations identified for both approaches (Benton and Elliot, 2016). Smith (1989) and Sadler (2009) have connected four practices that together could improve comparability: (1) selected exemplars that demonstrate how qualities may combine in a performance representative of a standard; (2) discussion among judges about their judgments in relation to the standard; (3) commentaries of the judgment decision, identifying the strengths and weaknesses of a performance and how these combined in the overall judgment; and (4) opportunities to judge and develop expertise.

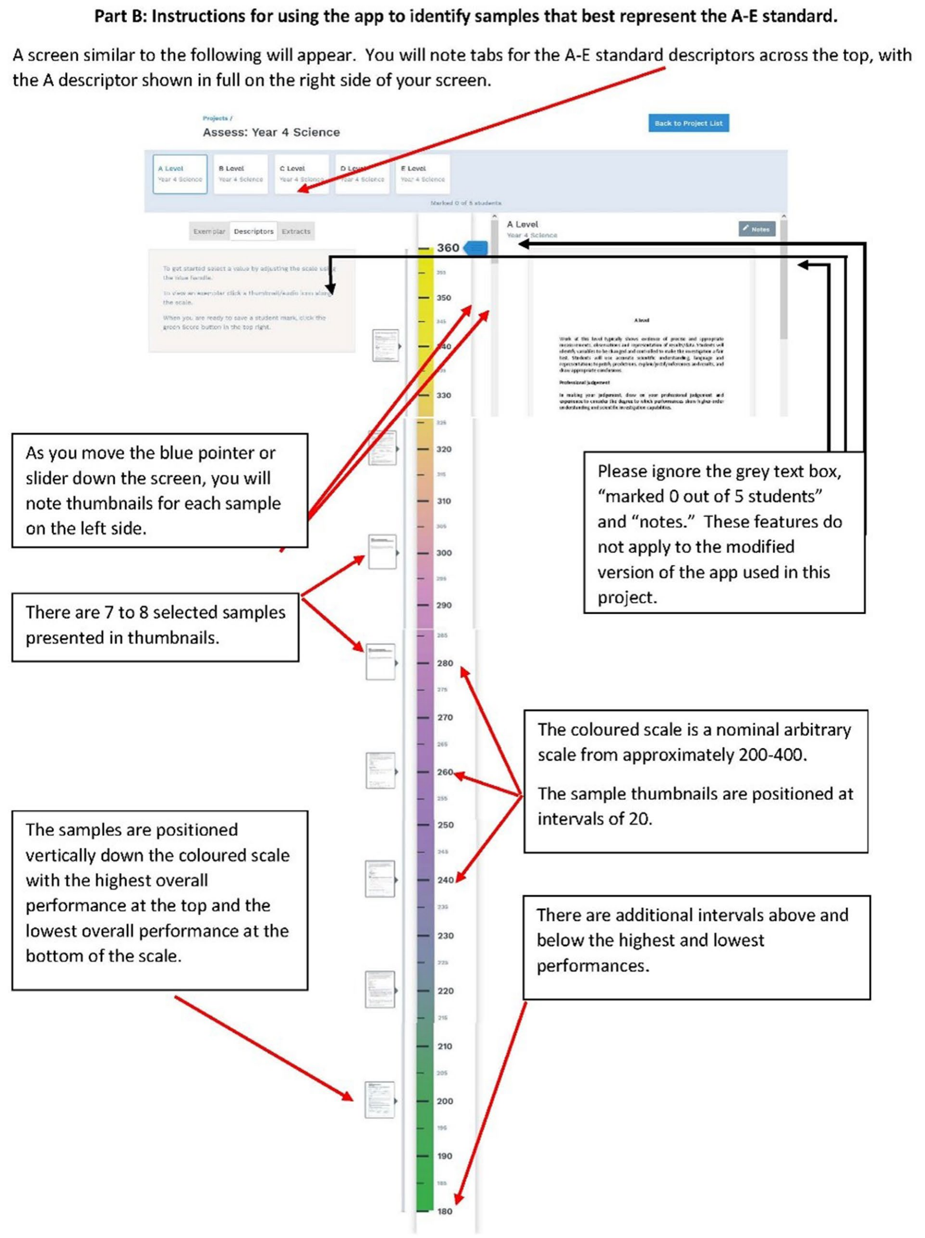
Given the fuzzy nature of standards, the scoring of complex performance assessments requires an approach that is different to those developed for dichotomous short-response and multiple-choice items with right and wrong answers. The focus in this paper addresses the first practice identified by Smith (1989) and Sadler (2009). This study does so by using quantitative and qualitative methods to select scaled samples with high scorer reliability that are representative of a standard.

Rationale

As stated, much of the literature on standard setting focuses on the application of procedures to tests comprising short-response items, which often use the Angoff method (Angoff, 1971), the modified Angoff method, and the Bookmark method (Lewis et al., 1996). One previous study that has attempted to apply an extended Angoff method to complex performances for standard setting was conducted by Hambleton and Plake (1995). In the study, panelists were required to specify the expected scores for just barely certifiable candidates on polytomously scored exercises in the context of professional teaching standards. The study produced mixed results; while there were high levels of agreement among the panelists, there was an indication that "panelists did not fully understand the implications of the extended Angoff procedure they had implemented" that "calls into question the validity of the resulting standard" (Hambleton and Plake, 1995, p. 53). Hambleton et al. (2000) later concluded that by design the Angoff and Bookmark methods are not readily applicable to extended performance; they noted a long-standing need for more suitable, novel, valid methods.

Prompted by the need for new approaches to set cut-points for performance assessment tasks, including those with multiple cut-points, Plake and Hambleton (2000) reported the use of a

[1] Smith (1989, 1995) now writes under the name of Wyatt-Smith.



Part B: Instructions for using the app to identify samples that best represent the A-E standard.

A screen similar to the following will appear. You will note tabs for the A-E standard descriptors across the top, with the A descriptor shown in full on the right side of your screen.

Projects /
Assess: Year 4 Science

To get started select a value by adjusting the scale using the blue handle.

To view an exemplar click a thumbnail/audio icon along the scale.

When you are ready to save a student mark, click the green Score button in the top right.

A level

Work at this level typically shows evidence of precise and appropriate measurements, observations and representation of results/data. Students will identify variables to be changed and controlled to make the investigation a fair test. Students will use accurate scientific understanding, language and representations to justify predictions, explain/justify inferences and results, and draw appropriate conclusions.

Professional judgement

In making your judgement, draw on your professional judgement and experience to consider the degree to which performances show higher-order understanding and scientific investigation capabilities.

As you move the blue pointer or slider down the screen, you will note thumbnails for each sample on the left side.

There are 7 to 8 selected samples presented in thumbnails.

The samples are positioned vertically down the coloured scale with the highest overall performance at the top and the lowest overall performance at the bottom of the scale.

Please ignore the grey text box, "marked 0 out of 5 students" and "notes." These features do not apply to the modified version of the app used in this project.

The coloured scale is a nominal arbitrary scale from approximately 200-400.

The sample thumbnails are positioned at intervals of 20.

There are additional intervals above and below the highest and lowest performances.

FIGURE 2
Instructions for teachers in Stage 2, standard setting, showing the different features of the app.

categorical standard-setting procedure applied to extended performance assessments, that is relevant to the current research. These authors investigated a judgment method in which "panelists review[ed] each of several student papers, sampled to present the full score continuum, and [made] categorical assignments regarding the performance levels represented by the quality of the student's work" (p.198). They concluded that further research was needed to explore the possibilities of panelists being able to "reconsider their classification decisions [and] to achieve… distinct differences in the quality of student papers assigned to each score category" (Plake and Hambleton, 2000, p.214). The current research builds on their work by implementing these types of refining strategies.

There has been relatively little research on standard setting for extended performances over the past two decades, with some exceptions particularly within the medical field. Kramer et al. (2003) applied a modified Angoff method to complex performances in the context of objective structured clinical examinations (OSCE) in which judges were required to rate the proportion of borderline candidates who would pass each station. Another study, conducted by Kaufman et al. (2000), required judges to rate each item on the checklist for each station which were averaged across judges and stations. They concluded that "a reasonably fair and accurate pass standard can be established using an Angoff procedure. However, a larger number of judges or stations would be required to obtain an acceptable level for the reliability of the pass/fail standard in the OSCE" (p. 270). Other studies have criticized the use of the Angoff method for performance-based clinical examinations; for example, Boursicot et al. (2006) found significant discrepancy in setting passing scores for OSCEs among different medical schools using the same standard setting method and the same stations.

Studies have also identified the importance of judge expertise to recognize quality performances based on shared understandings of performance standards. In the study by McKinley et al. (2005), panelists reviewed performances on a standardized patient examination of clinical skills to judge readiness to enter a medical education program. The authors concluded that a work-centered approach to standard setting was appropriate for performance-based assessments because judges could "use their expertise to determine which sets, or numbers, of actions were important in determining readiness" (p. 365). Roberts et al. (2017) compared two groups of panelists who generated their judgments under different administrative conditions and found they produced equivalent ratings. The study required panelists to make holistic judgments of qualified or not qualified, according to a performance standard, for performances on a Comprehensive Osteopathic Medical Licensing Examination that included standardized patient encounters.

There is a methodological parallel between the current research and that reported in Wyatt-Smith et al. (2020), on the one hand, and the Bookmark method, on the other. Specifically, in the Bookmark method, individual items are ordered by difficulty and presented to judges in the process of determining a standard. For example, the Bookmark method was used to re-set the cut-score for the Canadian Forces Firefighter Physical Fitness Maintenance Evaluation (Rogers et al., 2014). The authors concluded that the Bookmark method was applied successfully and that it can be used with an underlying continuum of a specific construct such as difficulty or time taken to complete a task.

In the current research and the GTPA context of Wyatt-Smith et al.'s (2020) study, pairwise comparison is a foundational step; extended performances are then presented, in order, to participants in the process of determining a standard. In the case of the Bookmark method, items are typically scaled using IRT models prior to standard setting. In the current research and in the GTPA context, performances are scaled using the Bradley-Terry-Luce (Bradley and Terry, 1952; Luce, 1959) model prior to standard setting. The present research extends the work in this field by focusing on multiple standards, rather than the single standard in the GTPA study. Having said this, although there are multiple standards per scale in this study, the same method can be applied to select a single standard based on extended performances, as was the case in the GTPA example (Wyatt-Smith et al., 2020).

The specific method introduced in this study is particularly useful where the standards set have an inherent ordering as is the case with *A* to *E* grade standards. Because judges can see the ordering of performances, they can readily select locations in a manner corresponding with the intended ordering. Having methodically ordered performances therefore makes the task for judges as straightforward as possible by presenting extended performances transparently in a validated order.

Design and development of the study

The design and development of the present study involved two stages: scaling based on paired comparisons and then standard setting. The first stage is similar to that reported in the studies in a writing context by Heldsinger and Humphry (2010, 2013) as well as Humphry and Heldsinger (2019, 2020). Note that, with the exception of the 2010 study, these previous studies comprised two stages but were concerned with assessment against calibrated exemplars and did not involve standard setting; in addition, their findings were limited to primary school contexts, while the present study includes middle school students' performances.

In Stage 1 of the present study, teachers compare students' written performances using the method of pairwise comparison in which they choose the better of two performances each time. The resulting data are analyzed to develop a scale on which the locations represent the quality of each performance. The scale is used to identify a subset of the performances that have been reliably judged. The subset is selected so that the performances are relatively evenly spaced on a highest to lowest scale, to act as calibrated exemplars in the following stage.

In Stage 2, the calibrated exemplars are presented in hierarchical order and judges are instructed to select the exemplar that best represents each of the *A* to *E* standard descriptors. This second stage relies on the evaluative expertise of the judge informed by broad *A* to *E* descriptors. Once each judge identifies the locations of standards, the variability of the scale locations selected as *A* to *E* standards for each year level is visually examined. This stage has similarities to a study by Wyatt-Smith et al. (2020) but differs in two respects: (1) the presentation of exemplars, and (2) the use of multiple standards, rather than a single standard.

Stage 1: pairwise comparison procedure

Selection of performances for stage 1

Performances of English narrative writing were collected from teachers in Queensland and Western Australia ($n = 625$). The tasks aligned with the description of the genre in the Australian Curriculum (Australian Curriculum, Assessment and Reporting Authority, 2022).

The large pool of 625 performances consisted of 465 performances from schools in Queensland and 160 performances from schools in Western Australia. The 465 Queensland performances were collected from Years 4, 6, and 8 from 19 government and independent schools, located in metropolitan and regional areas, with varying numbers of performances (between 5 and 99) obtained from each school. The 160 Western Australian performances were collected from Pre-Primary to Year 7 from seven schools[2]. Due to the limited number of schools in this study, it was not possible to employ a stratified random sample or other sampling design. Nevertheless, the schools were selected to reflect a range of values on the Index of Community Socio-educational Advantage (ICSEA) which is based on family background data (ICSEA 883–1,174, *M*: 1004, SD: 85).

From the original pool of 625 performances, 164 performances from Queensland and 115 performances from Western Australia (together totaling 279 performances) were selected for use in Stage 1. Performances were selected based on several criteria including approximately equal distribution of school awarded *A* to *E* grades,

[2] A detailed description of the performances, judges, and data analysis involving the WA performances from five of the schools is provided in Humphry and Heldsinger (2020). The present study included performances from two additional schools.

school size, school regions, and legibility. In addition, since the current study focuses on Years 4, 6, and 8, the 115 Western Australian performances were selected by first eliminating any performances with scores in the lowest quarter of the range, then selecting performances so there was an approximately uniform distribution of scores across the resultant range.

This resulted in a total of 279 written samples, collected from primary and secondary schools across two states in Australia, to be compared by judges in Stage 1 of this study.

Participating judges in stage 1

Thirty-one Queensland and 22 Western Australian teachers participated as judges in the pairwise comparisons of the performances from their respective states, which generated the total number of comparative judgments used in this first stage. Judges received training in the form of written and video instructions to make holistic judgments about students' writing skills, and to use the assessment and reporting software to make and record judgments (Figure 1). Judges compared performances on features of narrative writing including character and setting and conventions of writing. Figure 1 provides an extract from the written instructions on how to use the reporting software.

Pair generation for stage 1

A design for the pair generation was constructed in which each performance was compared with a number of other performances. In addition, the pairs were generated to satisfy specified criteria for accuracy of estimated scale locations. For example, while the pairs are allocated randomly to participating judges from the generated list, this allocation avoided duplicated comparisons by any given judge. That is, no pair was compared more than once by a single judge.

The standard error (SE) of the scale location of a performance is dependent on the number of comparisons as well as the relative locations of the performances it is compared against (Humphry and Heldsinger, 2020). For this reason, the maximum number of comparisons per performance is sought given the available time from judges and the anticipated number of comparisons that can be made in that time.

Pairwise comparison judgments in stage 1

Judges were given online access to specific pairs of written performances, as shown in Figure 1. They worked individually to compare pairs of performances based on holistic judgments as to which performance displayed more advanced writing skills. Each judge made between 41 and 121 comparative judgments, resulting in a total of 3,009 pairwise comparisons across the primary and secondary school performances.

These comparative judgments were combined with 3,532 comparative judgments of the original 160 Western Australian narrative performances (Humphry and Heldsinger, 2020), resulting in a total of 6,541 pairwise comparisons of 324 performances to be analyzed.

Analysis of pairwise data from stage 1

Data are analyzed using established techniques that are applied to paired comparisons. Thurstone (1927, 1959) established the *law of comparative judgment* which relates to a process of discrimination between pairs of objects based on an identified trait. This law is used in the design and principle for scale construction. Analysis of the pairwise data uses the number of times a performance is judged as better than other performances to estimate scale locations. The Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 1959) implements Thurstone's approach by substituting the logistic model for the normal distribution (Andrich, 1978).

In the present study, to scale performances, the judges' paired comparison data are analyzed using the BTL model. Specifically, the model is implemented using maximum likelihood estimation procedures. The strongest performance tends to have the highest proportion of 'better' judgments, though the estimation procedure takes into account the locations of the other performances against which samples are compared. Data-model-fit is evaluated using mean square standardized residuals, termed Outfit (Humphry and Bredemeyer, 2020). The person separation index is also computed to indicate the separation of locations relative to standard error, with values ranging from 0 to 1 and higher values indicating greater separation. The person separation index is analogous to Cronbach's alpha (Andrich, 1982).

Selection of exemplars

A subset of the performances was selected as calibrated exemplars. Performances were identified for inclusion in Stage 2 based on the statistical information about consistency of judgments combined with qualitative evaluation. There were three main criteria for selection: (1) successive performances from high to low are approximately equidistant on the scale; (2) paired comparison data for performances have acceptable fit to the model; and (3) discernible qualitative difference is evident between any two adjacent performances on the scale. The third criterion is performed for qualitative face validity verification.

The qualitative evaluation involved three subject experts reading each of the selected scripts to ensure ordering of samples and discernment between the quality levels of each performance. These experts met and discussed the ordering of each sample, suggesting alternate samples when there were disagreements, until an ordered set that met both quantitative and qualitative criteria was agreed upon.

From the available performances in each year level, a subset of eight written performances was selected as exemplars that most clearly and typically captured developmental features at given points on the scale. The aim was to select exemplars at equal intervals across the scale for each of the year groups.

Stage 2: standard setting exercise for face validity verification

Calibrated exemplars for use in stage 2

The eight calibrated exemplars for each year level, that were selected in the first stage of this study, were used for the standard setting verification exercise. The exemplars were ordered from highest to lowest to provide a clear and transparent process for judges to match with each of the broad *A* to *E* standard descriptors.

Participating judges in stage 2

A total of 28 teachers, 25 who participated as judges in the first stage of this study, were involved in the standard setting verification exercise. Eleven judges were involved in the standard setting of Year

4 performances, 10 for Year 6 performances, and 10 for Year 8 performances, with some judges participating in multiple year levels.

As in Stage 1, each teacher received training, in the form of written and video instructions, to make holistic judgments about students' writing skills, and to use the assessment and reporting app to make and record judgments. Figure 2 shows an extract from a page of the participant instructions that details the features of the standard setting app. The instructions for teachers to score a sample (exemplar) relative to a standard descriptor are presented in Figure 3 in a generalized form.

Standard setting procedure in stage 2

Judges were shown a set of exemplars on a scale and asked to select the score they considered to represent each standard. A linear transformation was applied to the scale obtained from the analysis of pairwise data so that the display range of the scale was 200 to 380 and increments of 5, making the range more readily interpretable for markers by avoiding negative numbers and decimals and any association with percentage (i.e., by not using a scale of 0 to 100). Judges were given online access to the eight calibrated exemplars for a year level, located on the scale, using the assessment and reporting software. The calibrated exemplars were positioned at intervals of 20, with the highest overall performance at the top of the scale and lowest overall performance at the bottom of the scale, with additional intervals above and below the highest and lowest performances.

Judges were asked to familiarize themselves with the exemplars on the scale, presented as thumbnail images, as well as the *A* to *E* standard descriptors. They then selected a sample on the scale that best represented each of the five descriptors in order, starting with the *A* standard. The selected samples correlated with a location on the scale for each standard level. The broad standard descriptions, derived from the Australian Curriculum, were presented on the screen adjacent to the exemplars. In some cases, judges identified the standard as being between two of the calibrated exemplars.

Judges worked individually and based their selection on holistic judgments. Each judge made five selections, for each of the five standard levels, for their chosen year level.

Results from stage 1 and stage 2

Pairwise comparisons

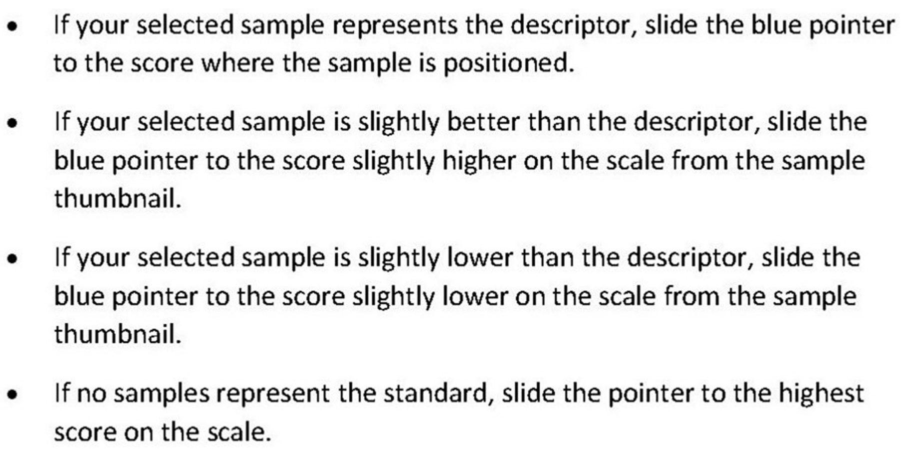
The analysis of the pairwise data identified one judge to be removed from the analysis, due to poor fit to the model (Outfit = 5.37, compared to a mean Outfit value of 0.70 for all 54 judges), resulting in a total of 53 judges. This removed 200 comparisons by that one judge. As a result of this removal, the person separation index increased from 0.951 to 0.962. This indicates a high level of internal consistency among judges, meaning a better ability to separate performances in terms of location estimates and a strong tendency for judgments to be consistent with the overall ordering of the performances.

Table 1 presents the summary statistics for the location estimates and Outfit values for all performances and judges in the joint analysis of Year 4, 6, and 8 performances. Higher Outfit values (over 1) indicate comparisons that are less consistent with the overall ordering. The Outfit values are centered about approximately zero though the mean Outfit value is lower than the expected value of approximately 1 because a relatively large number of performances ($n = 67$) have small Outfit values under 0.1. Overall, the data fit the BTL model well, with 29 performances of 324 having an Outfit value greater than 1.3.

Selection of performances displayed to judges

Three of the research team quantitatively examined performances. Quantitative selection criteria were (1) relatively equal intervals on the scale and (2) Outfit values which were not high. Any sample with an Outfit value over 1.5 was considered difficult to judge.

The qualitative evaluation involved three to four subject experts reading each of the selected scripts to ensure ordering. The main qualitative criterion was that the performances selected had discernible qualitative differences. The subject experts met and discussed the ordering of each sample, suggesting alternate samples when there were disagreements.



- If your selected sample represents the descriptor, slide the blue pointer to the score where the sample is positioned.
- If your selected sample is slightly better than the descriptor, slide the blue pointer to the score slightly higher on the scale from the sample thumbnail.
- If your selected sample is slightly lower than the descriptor, slide the blue pointer to the score slightly lower on the scale from the sample thumbnail.
- If no samples represent the standard, slide the pointer to the highest score on the scale.

FIGURE 3
Participant instructions for scoring samples (exemplars) relative to standard descriptors, during the standard setting process using the app.

TABLE 1 Summary statistics for location estimates and Outfit values for all performances and judges in the joint analysis of Year 4, 6 and 8 performances.

	N	Location				Outfit			
		M	(SD)	Min	Max	*M*	(SD)	Min	Max
Performances	324	0.00	(4.60)	−16.24	8.91	0.58	(0.59)	0.01	3.58
No extreme locations	284	−0.28	(4.44)	−15.85	7.78	0.65	(0.60)	0.01	3.58
Judges	53					0.63	(0.41)	0.02	2.24
No judgments >7*	53					0.59	(0.34)	0.02	2.23

*Individual comparisons with a standardized residual greater than 7 were removed.

TABLE 2 Summary of the equivalent scale locations, on the common scale, for the mean, median and SD of the scores indicated for each standard in Years 4, 6, and 8.

		N	Mean	Median	SD
Year 4	A	11	6.065	5.777	0.666
	B	11	3.884	4.365	1.130
	C	11	0.516	0.131	1.788
	D	11	−2.627	−2.691	1.309
	E	11	−4.552	−4.455	0.921
Year 6	A	9	6.600	6.888	0.809
	B	9	4.631	4.727	0.563
	C	9	1.894	1.702	1.334
	D	9	−1.372	−1.324	2.100
	E	9	−4.398	−3.917	1.914
Year 8	A	10	8.014	8.014	0.500
	B	10	5.528	4.806	1.032
	C	10	2.722	2.802	1.192
	D	10	0.637	−0.004	1.168
	E	10	−2.249	−2.009	1.183

The final ordered set met both quantitative and qualitative criteria.

Analysis of standard setting exercise

The results of the standard setting exercise are presented to illustrate the application of the method in a specific context and to make observations about outcomes of more general relevance in applying the method.

Table 2 shows the scale locations, on the common scale obtained from pairwise comparisons, for the mean and median scores selected for each standard by year level. Because a common scale was formed for Years 4, 6, and 8, the scale locations can be compared across the year levels. For example, the median scale location for the *A* standard is 6.888 in Year 6 and 5.777 in Year 4. Thus, the median is only somewhat higher in Year 6 compared to Year 4.

It can be seen that the median scores for *A* to *E* are in descending order. It can also be seen that the standard deviation (SD) of a given standard is typically substantially less than the difference between the median scale values of adjacent year levels.

Figures 4–6 summarize the distributions of the scores selected for each of the standards in Years 4, 6, and 8, respectively. The 25th to 75th percentile range is illustrated with a blue rectangle and the median as a black line within the blue area. The black lines at the top and bottom of the blue rectangle are the maximum and minimum scores respectively, excluding outliers if there are any. Outliers are shown as circles in Figures 4–6.

When presented to judges, a linear transformation was applied to the scale locations of the performances shown in Table 2 so that they lay within a range from 200 to 380. As stated, this is an arbitrary range chosen to avoid participants confusing scale locations with percentage scores and to avoid negative scale locations. Transformations of this kind are common in large-scale testing programs such as the OECD's Programme for International Student Assessment (PISA).

The main observation is that the medians for *A* to *E* are in descending order and the distributions are mostly, though not uniformly, tightly clustered around the median. Tightly clustered distributions indicate consistency among judges in the perception of the performances representing the relevant standard. For example, the scale values for the *A* standard in Year 4 are quite tightly clustered relative to the separation between the *A* and *B* medians.

Note that there is a relatively large range of scores selected for the *D* standard in Year 4, shown in Figure 4. There is smaller variation for the other standards, though with outliers present for the *B* and *C* standards.

Also note that there is significant variation for the *D* and *E* standards in Year 6, as shown in Figure 5. Generally, there is little variation for the other standards, but there are three outliers that are quite extreme.

Table 3 shows the mean and standard deviation of the scores selected by each judge across all standards. For example, for the first judge in the Year 4 exercise, the mean score across the *A, B, C, D,* and *E* standards is 288. The difference (Diff.) is that between the individual judge's mean score and the mean score given by all judges combined. This difference indicates whether the judge selected higher or lower scores overall. A positive difference suggests a general expectation of a higher standard of work to meet a given standard. The standard deviation indicates the spread of the scores selected by each judge across the standards. A higher standard deviation means the judge selected a larger spread of scores for the *A* to *E* standards. Within a single row of the table, it is not the same judge in each year level; the data are arranged this way in the table for ease of representation.

Limitations

The results of the project should be interpreted noting the following limitations. A relatively small number of judges (teachers) participated in the research, including pairwise comparison and standard setting stages. With a small sample, details of the results depend more on the

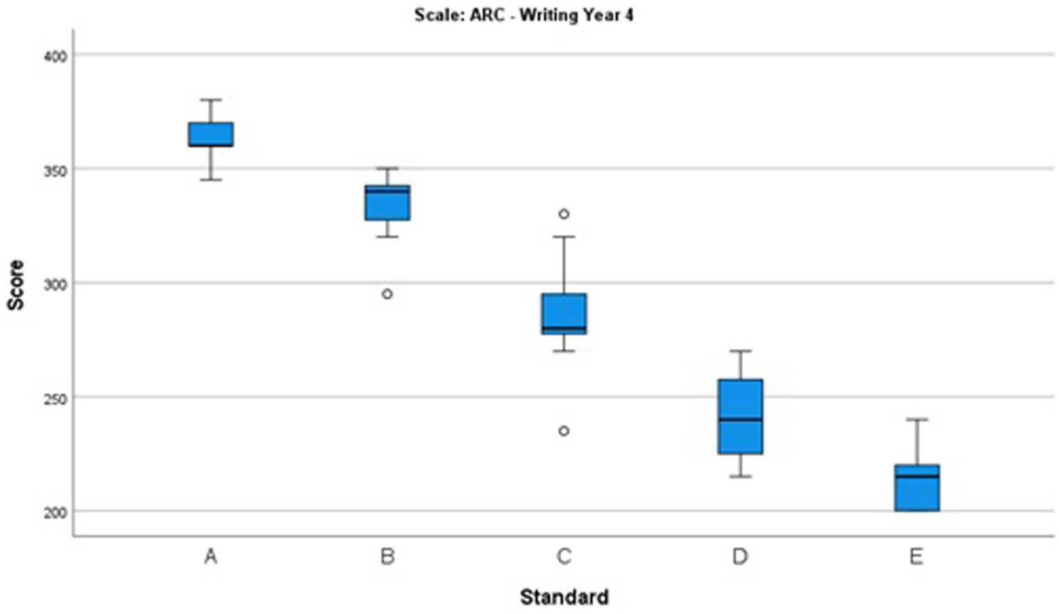


FIGURE 4
Box and whisker plot showing the distributions of the scores selected for each standard in Year 4.

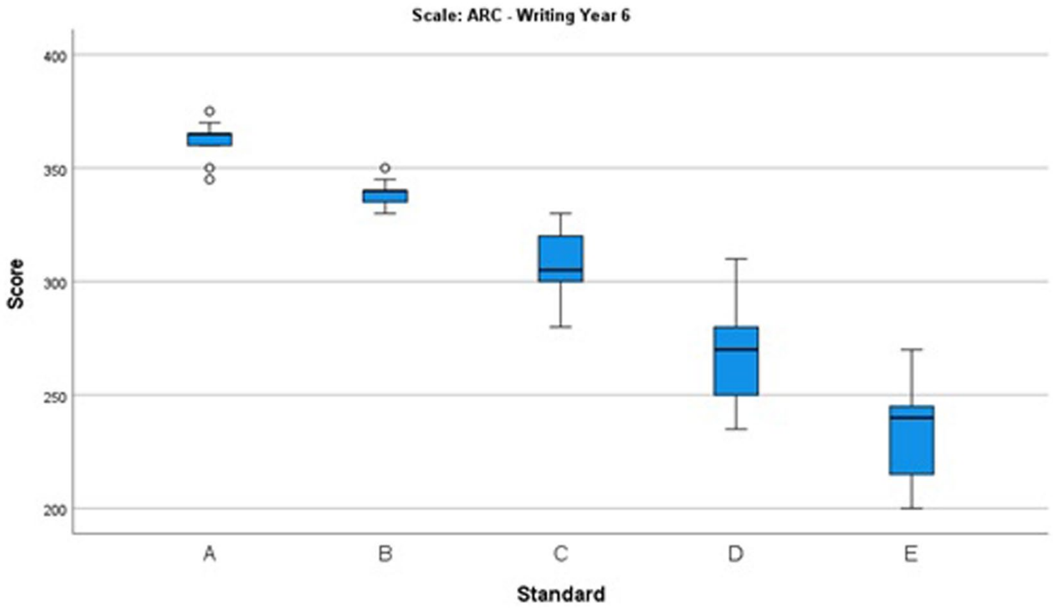


FIGURE 5
Box and whisker plot showing the distributions of the scores selected for each standard in Year 6.

specific schools and contexts of the participants than would be the case with a larger sample of teachers. The results are therefore indicative but cannot necessarily extrapolate to all teachers; for example, variation in where teachers place a *C* or *D* for a given year level may not be representative of all teachers. Where there was relatively large variation in the scores selected for a given standard, it is suggested that there was variability in judges' understanding of the quality indicators of that standard, which might be attributable to the limitations of a small sample of teachers as judges. The results also depend on the assessment tasks used and may or may not generalize to other tasks. In addition, the tasks might have been designed within a school and thus have not necessarily been through an extensive review process.

Discussion

As outlined in the introduction, the approaches to standard setting that are used most often are the Angoff and Bookmark methods, which are not readily applicable to extended performances (Hambleton et al., 2000). These methods also typically involve

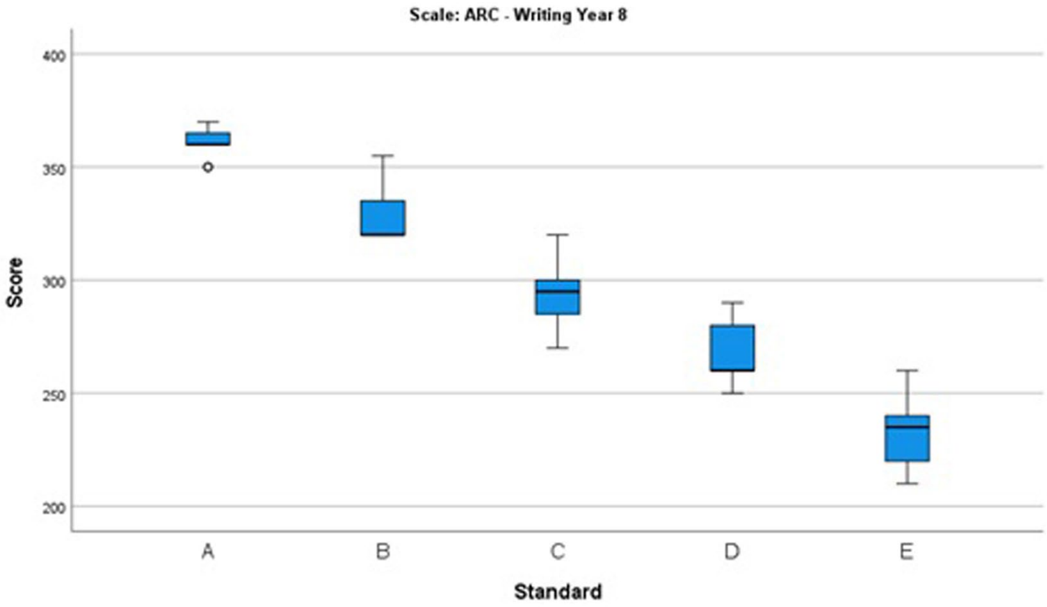


FIGURE 6
Box and whisker plot showing the distributions of the scores selected for each standard in Year 8.

TABLE 3 Summary of the range of scores selected by each judge across *A* to *E* standards in Years 4, 6, and 8.

	Year 4			Year 6			Year 8		
	Mean	SD	Diff.	Mean	SD	Diff.	Mean	SD	Diff.
Judge	288	61	1	296	69	−6	290	45	−7
	285	72	−2	290	66	−12	284	56	−13
	291	59	4	286	66	−16	295	60	−2
	308	52	21	299	41	−3	298	48	1
	296	57	9	323	35	21	312	54	15
	282	70	−5	301	43	−1	291	57	−6
	289	65	2	301	51	−1	294	53	−3
	302	68	15	312	50	10	296	48	−1
	285	57	−2	313	45	11	296	48	−1
	278	74	−9	234	27	−68*	310	41	13
	258	61	−29						
Mean	287			302			297		

Also indicated in the table is the difference between each judge's mean and the mean of all judges combined in each year level. Note: within a single row of the table, it is not the same judge in each year level; the data are arranged in this way for ease of representation. *Judge was removed and was not used in the calculation of the mean.

participants judging the probabilities of success on questions, which is often problematic because judges do not estimate absolute probabilities of success on individual questions very accurately (Lorge and Kruglov, 1953; Shepard, 1995; Impara and Plake, 1997, 1998; Humphry et al., 2014). This means that as a standard setting method, the novel SESS procedure has a key advantage in that it does not involve judging probabilities of success on tasks. Apart from the Angoff and Bookmark methods, other approaches to setting standards have been applied to a limited extent, including holistic judgments of whether a performance indicates the person is qualified or not qualified (e.g., Roberts et al., 2017).

Displaying a range of extended performances in order makes it more straightforward for judges to indicate what they perceive to be representative of a standard within the range. The method of paired comparisons has been shown to be effective for reliably scaling and ordering extended educational performances of various kinds (Humphry and Heldsinger, 2019). The approach outlined in this article exploits this for the purpose of standard setting by using paired comparisons as a preliminary first stage of a standard setting procedure that enables performances to be presented in order to judges. Participating judges may then select performances that they believe represent a given standard.

For the method to be effective, the order in which performances are presented needs to be valid. This article details quantitative and qualitative checks on the ordering of performances presented to the participating judges. Performances are selected based on scale locations, taking into account model fit. The fit index indicates whether a given performance is compared consistently, meaning it tends to be judged better than performances with a lower scale location and worse than performances with a higher scale location. Judgments are only expected to be consistent within the expectations of the BTL model and they need not be perfectly consistent to conform to these expectations.

Once performances are selected based on scale locations and model fit, they are also examined to check that a discernible qualitative difference is evident between adjacent performances on the scale. Qualitative examination of ordering of performances is key for effectiveness because standard setting is not straightforward if participants do not agree with the ordering of performances as they are presented.

Having obtained the mean, median, and distribution for each standard as perceived by the judges, shown in Table 2, it is possible to select performances that will be used to exemplify each standard for teachers in school settings. In the illustrative context, generally the performance nearest the median scale location was selected as the exemplar as the indicator of central tendency of the distribution of scale scores for the standard. For example, in Year 4 the *B* standard performance selected was the one nearest to the median scale score of 4.365 shown in Table 2.

For many of the standards there is relatively consistent selection of locations, as evident in relatively tight clustering in the box and whisker plots. However, in some cases there is larger variation of the locations selected for a given standard. Most notably, there is relatively large variation in the location of the *D* and *E* standards selected for both Years 4 and 6. There is also moderately large variation in the locations of the *C* standards selected. Where there is relatively large variation in the scale locations selected for a given standard, this indicates a variable understanding of the quality indicators of that standard.

Conclusion

This article describes and discusses the results of the Scaled Exemplar Standard Setting method (SESS), which has contributed new insights to the field in the context of extended performances, in this case in English narrative writing. Over the past two decades, there has been relatively little research on standard setting procedures for extended performances. Most of the literature has focused on the Angoff or Bookmark methods which are not suitable for extended performances; short-response testing; setting standards in practical examinations; or the use of a single standard. By contrast, the novel SESS method focuses on multiple standards and involves two stages. Stage 1 utilizes pairwise comparison of samples to scale performances and select exemplars, which are then evaluated quantitatively and qualitatively to ensure they are ordered consistently and validly. In Stage 2, standard setting is carried out by presenting the exemplars to judges in a validated order, so that each can make five selections in total that best represent each of the *A* to *E* standard descriptors.

The methods used, including the design and development of the project in each of the two stages, highlight some pertinent points. Using pairwise comparisons in the initial stage exploits the reliability of scaling and ordering performances that this method provides and allows exemplars to be presented to judges in order along the achievement scale. In addition, qualitative examination of ordering of performances, not just quantitative evaluation, is key because standard setting is not a straightforward process if participants do not agree with the ordering of the exemplar performances. Results are illustrative of an output of this novel method and give insight into the selection of calibrated exemplars as representative of each *A* to *E* standard. The mean and median scores selected for each standard, by year level, are displayed on a common scale and thus scale locations can be compared across year levels; median scores for *A* to *E* standards are in descending order; and the distributions of scores are predominantly tightly clustered around each median, indicating that judges were generally consistent in their selections for the majority of the standards. The main limitation of the study is that a relatively small number of judges participated in the research. With a small sample, the results obtained are indicative, but cannot necessarily be extrapolated to all teachers, schools or tasks. Where there was relatively large variation in the scores selected for a given standard, it is suggested that there was variability in judges' understanding of the quality indicators of that standard, which might also be attributable to the limitations of a small sample of judges.

Recommendations and implications for standard setting of extended performances confirm the utility of using paired comparisons in the first stage, to enable presentation of ordered performances to judges, and the importance of evaluating the validity of this ordering using quantitative and qualitative checks. The use of validated, ordered performances in the standard setting stage makes the task for judges as transparent and straightforward as possible. Overall, this research met its objective of focusing on the development of a standard setting method for extended performances, which would improve comparability of judgments, by employing quantitative and qualitative methods to select scaled exemplars with high scorer reliability that are representative of *A* to *E* standards.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This research was partially funded by an Australian Research Council Linkage Project (ARC LP180100046). We acknowledge the funding received for this project from the Australian Research

Council, Australian Catholic University, University of Western Australia and Partner Organisations: the Queensland Department of Education, Western Australian School Curriculum and Standards Authority, and Catholic Education Western Australia Limited.

Acknowledgments

Our sincere gratitude goes to the teachers in Queensland and Western Australia who contributed work samples, their time, and expertise in this research project. The authors also acknowledge the contribution to the outcomes of this research by project team members including Claire Wyatt-Smith, Chris DeLuca, David Andrich, and Terry Gallagher.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Appl. Psychol. Meas.* 2, 451–462. doi: 10.1177/014662167800200319

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR.20 index, and the Guttman scale response pattern. *Educ. Res. Perspect.* 9, 95–104.

Angoff, W. H. (1971). "Scales, norms, and equivalent scores" in *Educational measurement*. ed. R. L. Thorndike. *2nd* ed (Washington, DC: American Council on Education)

Australian Curriculum, Assessment and Reporting Authority (2022). *Australian Curriculum: F-10 curriculum*: English. Available at: https://v9.australiancurriculum.edu.au/f-10-curriculum/learning-areas/english/foundation-year_year-1_year-2_year-3_year-4_year-5_year-6_year-7_year-8_year-9_year-10?view=quick&detailed-content-descriptions=0&hide-ccp=0&hide-gc=0&side-by-side=1&strands-start-index=0&subjects-start-index=2

Benton, T., and Elliot, G. (2016). The reliability of setting grade boundaries using comparative judgement. *Res. Pap. Educ.* 31, 352–376. doi: 10.1080/02671522.2015.1027723

Bisson, M. J., Gilmore, C., Inglis, M., and Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *Int. J. Res. Undergrad. Math. Educ.* 2, 141–164. doi: 10.1007/s40753-016-0024-3

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.

Boursicot, K. A. M., Roberts, T. E., and Pell, G. (2006). Standard setting for clinical competence at graduation from medical school: a comparison of passing scores across five medical schools. *Adv. Health Sci. Educ.* 11, 173–183. doi: 10.1007/s10459-005-5291-8

Bradley, R. A., and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. *Method Paired Comp.* 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.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., and Mills, C. N. (2000). Setting performance standards on complex educational assessments. *Appl. Psychol. Meas.* 24, 355–366. doi: 10.1177/01466210022031804

Hambleton, R. K., and Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Appl. Meas. Educ.* 8, 41–55. doi: 10.1207/s15324818ame0801_4

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

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

Hudson, J., Bloxham, S., den Outer, B., and Price, M. (2017). Conceptual acrobatics: talking about assessment standards in the transparency era. *Stud. High. Educ.* 42, 1309–1323. doi: 10.1080/03075079.2015.1092130

Humphry, S., and Bredemeyer, K. (2020). The effect of interactions between item discrimination and item difficulty on fit statistics. *J. Appl. Meas.* 21, 379–399.

Humphry, S., 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., 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

Humphry, S., Heldsinger, S., and Andrich, D. (2014). Requiring a consistent unit of scale between the responses of students and judges in standard setting. *Appl. Meas. Educ.* 27, 1–18. doi: 10.1080/08957347.2014.859492

Humphry, S., 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

Impara, J. C., and Plake, B. S. (1997). Standard setting: an alternative approach. *J. Educ. Meas.* 34, 353–366. doi: 10.1111/j.1745-3984.1997.tb00523.x

Impara, J. C., and Plake, B. S. (1998). Teachers' ability to estimate item difficulty: a test of the assumptions in the Angoff standard setting method. *J. Educ. Meas.* 35, 69–81. doi: 10.1111/j.1745-3984.1998.tb00528.x

Jones, I., and Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educ. Stud. Math.* 89, 337–355. doi: 10.1007/s10649-015-9607-1

Jones, I., Swan, M., and Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *Int. J. Sci. Math. Educ.* 13, 151–177. doi: 10.1007/s10763-013-9497-6

Kaufman, D. M., Mann, K. V., Muijtjens, A. M. M., and van der Vleuten, C. P. M. (2000). A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Acad. Med.* 75, 267–271. doi: 10.1097/00001888-200003000-00018

Kramer, A., Muijtjens, A., Jansen, K., Düsman, H., Tan, L., and van der Vleuten, C. (2003). Comparison of a rational and an empirical standard setting procedure for an OSCE. *Med. Educ.* 37, 132–139. doi: 10.1046/j.1365-2923.2003.01429.x

Lewis, D. M., Mitzel, H. C., and Green, D. R. (1996). "Standard setting: a bookmark approach" in *IRT-based standard setting procedures utilizing behavioral anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on large-scale assessment*. ed. D. R. Green (Phoenix, AZ)

Lorge, I., and Kruglov, L. K. (1953). The improvement of estimates of test difficulty. *Educ. Psychol. Meas.* 13, 34–46. doi: 10.1177/001316445301300104

Luce, R. D. (1959). *Individual choice behavior*. New York, NY: Wiley.

McKinley, D. W., Boulet, J. R., and Hambleton, R. K. (2005). A work-centered approach for setting passing scores on performance-based assessments. *Eval. Health Prof.* 28, 349–369. doi: 10.1177/0163278705278282

Plake, B. S., and Hambleton, R. K. (2000). A standard-setting method designed for complex performance assessments: categorical assignments of student work. *Educ. Assess.* 6, 197–215. doi: 10.1207/S15326977EA0603_2

Potter, T., Englund, L., Charbonneau, J., MacLean, M. T., Newell, J., and Roll, I. (2017). Com-PAIR: a new online tool using adaptive comparative judgement to support learning with peer feedback. *Teach. Learn. Inq.* 5, 89–113. doi: 10.20343/teachlearninqu.5.2.8

Roberts, W. L., Boulet, J., and Sandella, J. (2017). Comparison study of judged clinical skills competence from standard setting ratings generated under different administration conditions. *Adv. Health Sci. Educ.* 22, 1279–1292. doi: 10.1007/s10459-017-9766-1

Rogers, W. T., Docherty, D., and Petersen, S. (2014). Establishment of performance standards and a cut-score for the Canadian forces firefighter physical fitness maintenance evaluation (FF PFME). *Ergonomics* 57, 1750–1759. doi: 10.1080/00140139.2014.943680

Sadler, R. (1987). Specifying and promulgating achievement standards. *Oxf. Rev. Educ.* 13, 191–209. doi: 10.1080/0305498870130207

Sadler, D. R. (2009). Grade integrity and the representation of academic achievement. *Stud. High. Educ.* 34, 807–826. doi: 10.1080/03075070802706553

Seery, N., Canty, D., and Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *Int. J. Technol. Des. Educ.* 22, 205–226. doi: 10.1007/s10798-011-9194-0

Shepard, L. A. (1995). "Implications for standard setting of the NAE evaluation of the NAEP achievement levels" in *Proceedings of the joint conference on standard setting for* *large scale assessments* (Washington, DC: National Assessment Governing Board and National Center for Educational Statistics), 143–160.

Smith, C. M. (1989). *A study of standards specifications in English. Master's thesis*. Brisbane, Australia: University of Queensland.

Smith, C. M. (1995). *Teachers' reading practices in the secondary school writing classroom: a reappraisal of the nature and function of pre-specified assessment criteria. Doctoral thesis*. Brisbane, Australia: University of Queensland.

Steedle, J. T., and Ferrara, S. (2016). Evaluating comparative judgement as an approach to essay scoring. *Appl. Meas. Educ.* 29, 211–223. doi: 10.1080/08957347.2016.1171769

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.* 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: The University of Chicago Press.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., and De Maeyer, S. (2016). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assess. Educ.* 26, 59–74. doi: 10.1080/0969594X.2016.1253542

Wyatt-Smith, C., Humphry, S., Adie, L., and Colbert, P. (2020). The application of pairwise comparisons to form scaled exemplars as a basis for setting and exemplifying standards in teacher education. *Assess. Educ.* 27, 65–86. doi: 10.1080/0969594X.2020.1712326