

## Abstract

National standardized assessment programs have increasingly included extended written performances, amplifying the need for reliable, valid and efficient methods of assessment. This article examines a two-stage method using comparative judgments and calibrated exemplars as a complement and alternative to existing methods of assessing writing. Written performances were taken from Australia's NAPLAN assessment, which included both narrative and persuasive performances from students aged 8 to 15. In Stage 1, assessors performed comparative judgments on 160 performances to form a scale of 36 calibrated exemplars. These comparative judgments showed a very high level of reliability and concurrent validity. In Stage 2, assessors scored 2380 new performances by matching them to the most similar calibrated exemplar. These matching judgments showed a generally high level of reliability and concurrent validity and were reasonably efficient after a familiarization period. Further research is suggested to enhance Stage 2 by simplifying the exemplar scale and scaffolding it with detailed descriptors. Overall, the findings support the use of the method in standardized writing assessment and its application to various learning areas.

### Applying a Thurstonian, two-stage method in the standardized assessment of writing

This article investigates the reliability, validity and efficiency of a two-stage method of writing assessment, which adapts the pioneering methodology of Thurstone (1928) and Thurstone and Chave (1929), with respect to written performances from Australia's large-scale standardized assessment program; the National Assessment Program - Literacy and Numeracy (NAPLAN). The two-stage process involves (a) the application of comparative judgments to a large set of written performances to establish a scale of calibrated exemplars, and (b) the subsequent use of these calibrated exemplars to assess new written performances through matching judgments. This two-stage method combines the benefits of comparative judgments in creating a reliable and valid 'ruler' of writing quality, where each calibrated exemplar represents a different level of quality to match with new performances, with the greater efficiency of matching judgments since comparative judgment designs are often criticized as highly time consuming. The two-stage method has been successfully applied to the assessment of writing in early childhood by classroom teachers with very high levels of reliability and concurrent validity observed (Heldsinger & Humphry, 2013). The current research extends the application of the method to the standardized assessment context, including both narrative and persuasive performances written by students across grades 3, 5, 7, and 9 (age range 8 to 15), and implements two criteria for the judgments in both stages of the process. The aim is to establish the reliability and validity of the two-stage method to explore its viability as a complement or alternative to current scoring methods in the standardized assessment of student writing.

### **Background**

As governments across the world increasingly invest in a range of large-scale assessment programs with the aim of enhancing accountability in their educational systems, the robustness of the standardized assessment of writing is becoming an increasingly

pertinent issue. Moreover, there has been a push for such assessments to develop beyond traditional multiple-choice formats to performance-based assessments that more closely align with the construct being investigated (Shermis, 2014). For example, under the Common Core State Standards initiative, which is presently transforming the educational assessment landscape in the United States, repeated, extended written performances are a central part of the formative and summative assessment of various content areas (Shermis, 2014). In the United Kingdom, extended written performances are a part of the various Key Stage assessments, and existing writing assessment practices have attracted a large amount of scrutiny and criticism (House of Commons Education Committee, 2017). Moreover, in the Australian context, extended written performances are already a core aspect of the annual National Assessment Program—Literacy and Numeracy (NAPLAN) assessments (Wyatt-Smith & Jackson, 2016).

The complexities associated with the standardized assessment of writing are intertwined with the complexities of the skill itself. Writing is a foundational aspect of students' literacy involving a complex set of cognitive processes and physical skills, which enables students to express, communicate, make connections and construct meaning in print (Hayes & Berninger, 2014; Mackenzie, Scull, & Munsie, 2013). Writing is also an integral part of schooling, as students beyond the age of eight may spend up to half of their classroom hours engaged in writing tasks across a range of learning areas, despite spending very minimal classroom time on learning and engaging in writing in the earlier grades of schooling (Brindle, Graham, Harris, & Hebert, 2016; McHale & Cermak, 1992). The minimal attention paid to the learning of writing in early schooling is noteworthy given that education systems across the world have increasingly expressed concern regarding declines in students' writing achievement (Mackenzie et al., 2013; Wyatt-Smith & Jackson, 2016), which has also motivated the broadening of standardized assessments beyond the traditional tests of

mathematics and reading (Mackenzie et al., 2013). With the increase in the volume of written performances in high-stakes standardized assessments, there is a growing need to develop methods that optimize the validity and reliability of writing assessment, as well as the efficiency with which they are implemented by assessors and teachers.

Standardized writing assessment, to date, has been dominated by analytic, rubric-based methods where written performances are rated on multiple criteria intended to cover the scope of the writing construct (Spandel, 2005). These rubric-based methods are also used across a range of learning areas throughout the school years and across a range of disciplines in higher education (Heldsinger & Humphry, 2013; Rezaei & Lovorn, 2010; Sadler, 2009; Tierney & Simon, 2004). It has been argued that rubric marking carries a number of advantages such as providing criterion-level diagnostic information, bridging formative and summative forms of assessment, and reducing subjectivity with high levels of reliability between raters (Brookhart & Chen, 2014; Jonsson & Svingby, 2007; Reddy & Andrade, 2010). However, a number of authors have argued that the validity and reliability of the rubric approach have not been adequately empirically studied (Heldsinger & Humphry, 2013; Humphry & Heldsinger, 2014; Rezaei & Lovorn, 2010).

Specifically, rubric-based assessment has been shown to be affected by numerous rating tendencies, including rater leniency, central tendency, restriction of range and the halo effect (Myford & Wolfe, 2003). Humphry and Heldsinger (2014) provided evidence that the common, grid-like structure of rubrics, where each criterion has a common number of categories, induces violations of local independence across ratings, which undermines validity by limiting construct-relevant variation in scores. Rezaei and Lovorn (2010) showed that if raters were not well trained in employing a rubric effectively, their assessments were biased toward the more mechanical features of writing. This latter point is particularly pertinent to large-scale, high-stakes standardized writing assessments, as effective assessor

training requires time and resources. In response to some of these limitations, a number of authors have proposed the application of comparative judgments as an alternative approach to writing assessment, and to performance assessments more generally (Bramley, Bell, & Pollitt, 1998; Heldsinger & Humphry, 2010, 2013; Pollitt, 2012; Steedle & Ferrara, 2016; Tarricone & Newhouse, 2016).

### **The Comparative Judgment and Two-Stage Approaches to Writing Assessment**

The theoretical background and practical applications of the comparative judgment approach is overviewed in detail elsewhere (Andrich, 1978; Bond & Fox, 2012; Bramley et al., 1998; Heldsinger & Humphry, 2010; Thurstone, 1927). The approach was proposed by Thurstone (1927) as a method to scale the perceived property of a set of stimuli (e.g., their perceived mass) by statistically modelling comparisons of the property of the stimuli in pairs. In the present context, a set of writing performances may be scaled in terms of their perceived quality by comparing them in pairs and inferring the scale locations from the proportion of judgments in favor of each performance versus each other performance. While there have been few educational applications of the comparative judgment approach compared to other methods of writing assessment, Heldsinger and Humphry (2010, 2013) demonstrated that it could be used to order performances in a highly reliable and valid manner across a range of learning areas.

A distinctive advantage of comparative judgments over rubric ratings is that teachers directly compare one performance against another so that there is reduced possibility for assessor harshness or lenience to affect judgments (Andrich, 1978; Heldsinger & Humphry, 2010). Similarly, because there is no rating required, comparative judgments are not subject to the biasing effects of different rater tendencies, or the response dependencies that are induced by structural features of the design of rubrics (Humphry & Heldsinger, 2014).

Heldsinger and Humphry (2010) also showed that the comparative judgment approach could be efficiently implemented with a very high level of reliability across different content areas in a classroom context in the absence of the extensive training processes often employed in large-scale testing programs. The primary disadvantage of comparative judgments, however, is that they can be time-consuming and therefore cognitively intensive for the judges (Bramley et al., 1998; Pollitt, 2012; Steedle & Ferrara, 2016).

To overcome this disadvantage, Heldsinger and Humphry (2013), drawing upon the work of Thurstone (1928) and Thurstone and Chave (1929), implemented a two-stage process designed to exploit the above advantages of comparative judgments while attaining greater efficiency. As briefly described above, the first stage of the process is to create a scale of exemplars by pairwise comparisons of written performances. Subsequently, teachers assess new performances using this scale of exemplars by matching the quality of the novel written performance with the most similar calibrated exemplar on the pre-existing scale. Heldsinger and Humphry (2013) applied this two-stage method in the assessment of narrative writing for early childhood and, despite very minimal training, found very high levels of inter-judge reliability (the average was .923). The authors also cross-referenced assessments using the two-stage method with rubric ratings and found a high level of concurrent validity ( $r = .895$ ). Finally, Heldsinger and Humphry (2013) observed that the method was a reasonably efficient way of obtaining highly reliable judgments with minimal training, as teachers on average were able to assess approximately 20 written performances per hour.

The present study aims to extend the findings of Heldsinger and Humphry (2013) in several important respects. First, Heldsinger and Humphry (2013) applied the method in a classroom context, with an early childhood student cohort and only a single writing genre. The current research applies the method to performances taken from NAPLAN over two calendar years, including performances by students aged 8 to 15, and both narrative and

persuasive genres. Moreover, Heldsinger and Humphry (2013) implemented the method using a single holistic judgment. The present study extends this research by implementing two criteria, writing conventions and authorial choices, in both stages of the two-stage process. The inclusion of separate judgments for these two criteria was based on feedback from teachers that, in some instances, it was difficult to reconcile these two broad aspects of the written performances to make a holistic comparison. The study therefore also extends the existing research on the use of comparative and matching judgments for writing assessment, which have generally relied upon holistic judgments, by examining the two-stage method using two criterion judgments. These two judgments, in turn, provide more information about the quality of the written performance. The method is investigated in terms of reliability, concurrent validity and efficiency, to ascertain whether it constitutes a viable complement or alternative to existing methods in the standardized assessment of student writing.

### **The study**

#### **Procedure – Overview**

In the first stage of the method, both narrative and persuasive writing performances were compared by experienced assessors using the method of pairwise comparison (Thurstone, 1927). Comparative judgments were made with respect to two broad criteria: (i) authorial choices, which includes features of writing where the writer is free to make choices including subject matter, language choices, development of tone, style, voice and reader-writer relationship; and (ii) conventions, where the writer is expected to largely follow rules, including spelling, punctuation, correct sentence formation, and clarity of referencing. Assessors were instructed to make an on-balance comparison for each broad criterion judgment, rather than favoring any particular aspect of the criterion, which is consistent with previous research using comparative judgments for writing assessment (Heldsinger & Humphry, 2010, 2013). Only these concise criteria were provided to orient assessors to the

key features of performances that they need to focus on when judging which in a given pair is better because, unlike rubric-style approaches, this process does not necessitate the description of gradations or levels for each criterion. Nonetheless, while concise criteria were used in the present context, all assessors involved in the study were able to draw from their familiarity with the writing marking guides for NAPLAN (Australian Curriculum, Assessment and Reporting Authority (ACARA), 2010, 2013).

The NAPLAN Narrative and Persuasive writing marking guides include ten marking criteria and only differ with respect to a single criterion (ACARA, 2010, 2013). The first six criteria of each guide are subsumed within the broad ‘authorial choices’ criterion of the present study, including ‘Audience’, ‘Text structure’, ‘Ideas’, ‘Character and setting’ (in the Narrative guide only), ‘Persuasive devices’ (in the Persuasive guide only), ‘Vocabulary’, and ‘Cohesion’. The last four criteria of each guide are subsumed within the broad ‘conventions’ criterion of the present study, including ‘Paragraphing’, ‘Sentence structure’, ‘Punctuation’, and ‘Spelling’. For further information see ACARA (2010, 2013).

The comparative judgment data were analyzed using the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959) to develop authorial choices, writing conventions and combined criteria scales on which the ordering of location estimates represent the judged ordering of the quality of the written performances on the respective criteria and overall. The combined criteria scale estimates were obtained by combining all comparative judgments in the BTL model analysis, irrespective of which criterion was used for the judgment. Moreover, performances across all grades and across both genres were simultaneously calibrated, which is consistent with the operational NAPLAN procedures where students across the grades in a calendar year respond to a common prompt and performances are scored on a common NAPLAN writing scale, irrespective of the writing genre. Humphry and McGrane (2015) also show how comparative judgments are operationally used to



longitudinally equate the NAPLAN writing scale in the absence of a common-person or common-item equating design. After calibration was complete, a subset of performances representing the full range of performances from the combined criteria scale was then selected for use as *calibrated exemplars*.

In Stage 2, ACARA engaged experienced assessors to score a new pool of narrative and persuasive written performances from two years of the NAPLAN writing assessment. The assessors scored each of the new written performances by judging which calibrated exemplar from Stage 1 was most alike the new performance, and the latter was assigned the scale estimate of the matched calibrated exemplar. Each assessor made a matching judgment of each performance separately for the two criteria and they were instructed to make an on-balance judgment for each broad criterion.

The Stage 2 data were examined to establish the inter-judge reliability of the matching judgments in terms of the two criteria and a combined scale. In addition, the concurrent validity of the matching data was assessed by comparing the estimates with a traditional rubric score assigned to the same performances by a different set of highly experienced assessors using the above described NAPLAN writing marking guides (ACARA 2010, 2013). Finally, the efficiency of the methodology was examined by inspecting the duration of the matching judgments over the full set of judgments by the assessors.

### **Stage 1 – Calibration of the Exemplars**

**Stage 1 method.** The results of Stage 1 are summarized as the Stage 2 results are the focus of this article.

**Participants.** The student performances were sampled from two years (2011 and 2012) of NAPLAN. The NAPLAN writing tests are administered under standardized conditions according to protocols. Time is allowed to each student in each grade level for

planning (five minutes), writing (30 minutes) and editing (five minutes). In a given calendar year, every student from each grade level received the same writing prompt/topic.

A total of 3000 performances were obtained from a clustered random sample across grades 3, 5, 7 and 9. From this larger sample, a subsample of 160 performances, including 80 narrative performances from the 2011 NAPLAN writing assessment and 80 persuasive performances from the 2012 NAPLAN writing assessment, was obtained for the comparative judgment stage of the study. This subsample was selected to have an approximately uniform distribution with between 1 and 8 performances for each score in the range from 0-48 (min-max possible score) on the NAPLAN rubric. The subsample was also drawn to have approximately equal numbers of students from each grade level (grade 3  $n = 43$ ; grade 5  $n = 38$ ; grade 7  $n = 38$ ; grade 9  $n = 41$ ).

***Participating assessors.*** Twenty-one expert assessors drawn from a pool of experienced NAPLAN assessors in one Australian State were involved as judges in Stage 1. Assessors made independent judgments either in a central venue or at other sites.

***Comparative judgment procedure.*** The number of possible paired comparisons in which each performance is compared with every other, is  $160(160-1)/2=12720$ . Given it is not possible to efficiently make this number of comparisons, a design was constructed so that across the 21 assessors, each performance was compared with an average of 40 other performances, giving an average of 80 judgments across the two criteria. This number of comparisons was used to attain sufficiently small standard errors of measurement for selecting ordered, calibrated exemplars to be used in Stage 2. The specific pairs compared by each individual assessor were randomly selected from the list of all pairs.

***Data analysis.*** The authorial choices, conventions and combined scales were produced by analyzing the criterion-level and combined comparative judgments using the BTL model. This model is essentially the same as Thurstone's (1927) Case V of the Law of

Comparative Judgment, but the logistic function is substituted for the cumulative normal function. The BTL is also mathematically equivalent to the conditional form of the Rasch model when the item parameter has been eliminated (Andrich, 1978). The BTL model is

$$p_{ij} = \Pr(i > j) = \frac{\exp(\theta_i - \theta_j)}{1 + \exp(\theta_i - \theta_j)} \quad (1)$$

where, in the case of written performance judgments,  $\theta_i$  and  $\theta_j$  in Equation 1 represent the level of writing quality instantiated by the performances  $i$  and  $j$ , and  $p_{ij}$  is the probability that performance  $i$  is judged better than performance  $j$ . In general terms, written performances with a low proportion of favorable judgments obtain more negative estimates, performances with a high proportion of favorable judgments obtain more positive estimates, and performances that have a similar proportion of favorable judgments receive similar estimates. The BTL model parameters were estimated using the PairWise software (Holme & Humphry, 2008), which implements maximum likelihood estimation and constrains the mean of the estimates to zero. The PairWise software also provides a Person Separation Index (PSI), which is analogous to Cronbach's (1951) alpha coefficient (Andrich, 1988) and provides an index of the internal consistency of the judgments as a whole. The PSI is defined as the ratio of true variance to observed variance in the writing quality estimates,

$$PSI = \frac{\text{var}[\hat{\theta}] - MSE}{\text{var}[\hat{\theta}]} \quad (2)$$

where

$$MSE = \sum_n \sigma_n^2$$

is the mean square error of the writing quality estimates from the BTL model. The PSI ranges in value from 0 to 1, with higher values indicating greater reliability in the comparative judgments.

### **Stage 1 results.**

***Comparative judgment reliability.*** The PSI was .971 for the authorial choices scale, .967 for the conventions scale, and .982 for the combined scale. These results indicated that all judgments had a very high level of reliability. The PSIs for the specific criterion and performance genre combinations are provided in Table 1. These similarly showed a very high level of reliability, ranging from .971 to .983 for the narrative performances across the criteria, and .962 to .980 for the persuasive performances across the criteria. The PSIs calculated by grade level were similarly very high, ranging from .939 to .968 for grade 3, .952 to .977 for grade 5, .960 to .978 for grade 7, and .960 to .976 for grade 9, across the criteria. These were not further differentiated into performance genre, as the sample sizes were not sufficient to obtain robust estimates.

***Correlations between judgment criteria.*** The correlation between the estimates obtained from the authorial choices and conventions criteria comparative judgments for each of the performances was .964. This correlation was very similar for both the narrative performances (.958) and the persuasive performances (.973). Moreover, it was also very similar for grade 3 (.948), grade 5 (.969), grade 7 (.960), and grade 9 (.936). These findings provide justification for combining criteria in an analogous manner to combining criteria in rubrics with mutually correlated scores. Moreover, the magnitude of these correlations suggests that the judgments for these two criteria may not be discriminating between different aspects of writing quality. Nonetheless, given that assessors from a previous application of the two-stage method indicated that, for select performances, it was difficult to reconcile these aspects to make a holistic comparison, the two criteria were carried over to Stage

**2. *Concurrent validity with rubric scores.*** To establish the concurrent validity of the comparative judgment procedure, the three comparative judgment criteria scales were correlated with the rubric scores for the 160 performances. This correlation was .944 for the authorial choices scale, .944 for the conventions scale, and .951 for the combined scale. Further, these correlations were essentially identical when examined for just the narrative performances (.946, .943, & .953), and just the persuasive performances (.942, .946, & .948). These results provide strong evidence that the concurrent validity was not moderated by the type of writing performance.

## **Stage 2 – Using the Matching Procedure with Calibrated Exemplars to Assess Writing**

**Stage 2 method.** In Stage 2, experienced assessors scored a new set of narrative and persuasive writing performances using 36 calibrated exemplars from Stage 1. This number of exemplars was selected to cover the full range of NAPLAN writing performance quality across the four grades and to provide a similar level of granularity in scores to the NAPLAN analytic rubrics. Sufficient comparative judgments were performed in Stage 1 to ensure with reasonable certainty that the 36 exemplars were correctly ordered. Moreover, the exemplars were carefully selected to ensure that they were consistent in their ordering for both the authorial choices and conventions criteria, as determined by their proximity to the regression line between the criteria estimates in Stage 1. Finally, a content expert involved in the NAPLAN operations reviewed the exemplars to ensure that they were free of any anomalous content that may affect the assessors' perception of their relative ordering.

A linear transformation was carried out on the scale of exemplars displayed to assessors so that their scores ranged from 50 to 95, which was a more familiar score range to these assessors than the logit scale estimates from Stage 1<sup>1</sup>. The performance scale was displayed graphically as shown in Figure 1 with reference exemplars to the left of a scale and

---

<sup>1</sup> The lowest score of 50 was not intended to indicate a 'pass' performance, but rather as a more intuitive lower bound for assessors than the negative values of the logit scale.

the performances to be assessed on the right-hand side of the display area. Assessors were asked to match each performance assessed to a reference exemplar on the scale, as described in more detail in the procedure section to follow.

As described above, a key difference from Heldsinger and Humphry (2013) in the current study was that matching judgments were made using both the authorial choices and conventions criteria from Stage 1, rather than a holistic judgment of performance quality. Thus, each of the new narrative and persuasive written performances in Stage 2 was assessed by matching the performance qualities to the most similar calibrated exemplar from Stage 1 with respect to the specific broad assessment criterion. Each performance was assigned the logit scale estimate of the calibrated exemplar with which it was matched for that criterion. In addition, a combined scale estimate was calculated for each performance in Stage 2 by averaging their matched scale estimates across the authorial choices and conventions criteria. These matched logit scale estimates were used in the calculation of all inter-judge reliability correlations, which are explained further in the Data Analysis subsection below.

***Collection of the pool of writing performances for assessment at Stage 2.*** The new pool of narrative and persuasive writing performances was drawn from the larger sample of approximately 3,000 writing performances from which the 160 Stage 1 performances were also drawn. In total, 2380 new performances were selected, including 586 from grade 3, 532 from grade 5, 679 from grade 7, and 583 from grade 9.

***Participating assessors.*** Twenty-one experienced assessors participated in the Stage 2 assessment. They were drawn from a pool of experienced NAPLAN assessors in a different Australian State from Stage 1. Assessors made independent judgments either in a central venue or at other sites.

***Procedures.*** Assessors were provided with brief instructions regarding the assessment task to match new written performances to their most similar exemplars, and were instructed

to make an on-balance judgment for the two broad criteria. No further training was provided. To complete the task, assessors were digitally presented with 36 exemplars displayed adjacent to a vertical scale in customized software, as shown in simplified form in Figure 1. Assessors were also provided with a hardcopy booklet containing the 36 exemplars and a response sheet. Each of the written performances to be matched was presented on the right-hand side of the screen. Performances were presented in a random sequence, with each assessor making a total of 346 matching judgments (173 each for the authorial choices and conventions criteria). Once assessors had determined the reference exemplar *X* that was closest in quality to the assessed performance *Y* for that criterion, they assigned that scale score of *X* to the assessed performance by dragging an icon into a box at the appropriate scale location adjacent to that exemplar. Assessors first matched their allocated writing performances based on the authorial choices criterion and then for the conventions criterion. The duration of the judgment was also recorded for each matching judgment on each criterion.

**Data analysis.** Similar to Heldsinger and Humphry (2013) and Gyagenda and Engelhard (2009), the inter-judge reliability for all of the 210 possible pairs of assessors was computed as the mean of the inter-judge correlations in their matching judgment estimates for the commonly judged performances. These were calculated for the authorial choices, conventions and combined criteria, as well as for the individual grade levels. As an additional indicator of inter-judge reliability, the correlations between the matched estimates of each assessor and the mean estimates for all assessors (excluding the assessor being compared) were calculated for the authorial choices, conventions and combined criteria. Moreover, the correlation between the authorial choices and conventions judgment criteria estimates was calculated and the median durations of matching judgment time were examined to explore the method's efficiency.

**Stage 2 results.**

***Inter-judge reliability.*** The average inter-judge correlation between each of the 210 unique pairs of assessors was high for each of the judgment tasks, indicating good inter-judge reliability for the matching judgments. The mean correlation was .799 (range .585 to .919) for the authorial choices criterion, .781 (range .606 to .928) for the conventions criterion, and .858 (range .702 to .932) for the combined criteria. These mean inter-judge correlations were similar for just narrative and just persuasive performances, although larger for the latter (see Table 1). Moreover, they were similar for three of the individual grade level performances, ranging from .746 to .838 for grade 3, .719 to .822 for grade 5, and .689 to .820 for grade 9, across the criteria. The mean inter-judge correlations were lower for grade 7, ranging from .569 to .672 across the criteria, although the variability in the estimates was much smaller for this grade (*SD* range from 1.442 to 1.495) compared to the other grades in the sample (*SD* range from 1.742 to 2.279). Similar to Stage 1, the grade level correlations were not further differentiated into performance genre due to insufficient sample sizes.

The correlations between each assessor's matching estimate and the average estimate of all other assessors similarly indicated a high level of inter-judge reliability for each of the matching judgment tasks. The mean correlation was .893 (range from .798 to .955) for the authorial choices criterion, .888 (range from .805 to .956) for the conventions criterion, and .928 (range from .847 to .953) for the combined scale. These mean correlations were similar when examining the judgments for just the narrative performances (.857, .842, & .901) and the persuasive performances (.928, .919, & .950).

***Correlations between judgment criteria.*** The correlation between the average estimates obtained from the authorial choices and conventions criteria judgments was .969, where the average was taken across all assessors. This correlation was similarly very high for the subsets of narrative performances (.949) and the persuasive performances (.980). This



finding provided evidence that the strong correspondence between judgment criteria was not substantially moderated by performance genre.

***Concurrent validity with rubric scores.*** To establish the concurrent validity of this method using calibrated exemplars, the estimates obtained by the matching judgments were then correlated with scores obtained from the NAPLAN analytic rubric (ACARA, 2010, 2013) for the same performance scripts.

***Procedure.*** Six experienced NAPLAN assessors double-marked the full sample of 2,380 performances using these rubrics. This marking exercise was undertaken separately from Stage 1 and 2 by a separate group of assessors. The assessments were made according to the procedures for marking NAPLAN writing performances set out in the above described NAPLAN writing marking guides (ACARA 2010, 2013). This rubric assessment provided a total score, authorial-choices sub-score for the six criteria consistent with this broad criterion, and conventions sub-score for the four criteria consistent with this broad criterion. Similar to the findings for the comparative judgment procedure and matching procedure estimates, these NAPLAN rubric sub-scores were highly inter-correlated, with an overall correlation of .929 across all performances, .947 for the persuasive performances, and .907 for the narrative performances.

The NAPLAN rubric has consistently had a PSI of between 0.95 and 0.96 over the course of the testing program, which shows a very high, historical level of reliability. The inter-rater reliability for this set of written performances was very high for the total rubric scores with an inter-rater correlation of .894 for the narrative performances and .922 for the persuasive performances. However, these correlations may be somewhat inflated relative to the general level of reliability in NAPLAN rubric marking across all States, as these six raters were some of the most experienced raters from a single marking center, and had therefore undergone training on the rubric together over a number of years. Given that all raters were

highly experienced and showed a high level of inter-rater reliability, the final rubric score and sub-scores used for correlations with the matched estimates were the average scores of the two raters for each performance.

*Concurrent validity results.* To first examine the concurrent validity of the matched estimates, these average NAPLAN rubric scores and sub-scores were correlated with the average matched estimates (taken across both criteria) for the 37 performances that were matched by all 21 assessors, as these estimates are based on multiple-observations and so should be more reliable. These correlations showed a very high level of concurrent validity between the matched estimates and NAPLAN rubric scores. The correlation with the total rubric score was .925 for the average authorial choices matched estimate, .926 for the average conventions matched estimate, and .933 for the average combined matched estimate. These correlations were similar for just the narrative performances (.917, .908, & .924) and just the persuasive performances (.942, .951, & .952), although they were somewhat larger for the latter.

Next, the individual matched estimates of all of the performances were correlated with their rubric score and sub-scores, and showed a similarly high level of concurrent validity with correlations ranging from .819 to .900 for the combined and individual performance genres (see Table 1).

*Efficiency of matching judgments.* The computerised administration of Stage 2 enabled the recording of judgment duration for each assessor across their 346 judgments. To overcome the influence of outliers, the median duration for all assessors was examined. The judgments showed a consistent downward trend in average duration from the beginning to the end of task. The median duration commenced at 16.36 minutes for the first matching judgment. By a third of the way through the task, the median duration fell to 6.68 minutes per judgment. By the final judgment, the median duration was 3.77 minutes compared to an

overall median duration of 6.03 minutes for all of the matching judgments across the entire task.

## **Discussion**

### **Discussion of the Stage 1 Results**

Similar to the previous findings of Heldsinger and Humphry (2010, 2013), the development of the scale of exemplars by comparative judgments in the Stage 1 calibration showed a very high level of reliability. The separation indices were all very high and similar for the different genres, criteria and grades. These high separation indices may be attributable in part to the number of judges, as explained by Heldsinger and Humphry (2013), and so they are not directly comparable to an index of a rubric marked by a single judge. Nonetheless, the indices show that the assessors were very consistent in their interpretation and assessment of the written performances, and this consistency was not moderated by genre, criterion or grade. The very high correlation between the two criterion judgments in Stage 1 also provided strong evidence that it was appropriate to form a single scale of calibrated exemplars for the matching judgments in the second stage of the study.

### **Discussion of the Stage 2 Results**

The aim of stage two was to ascertain the reliability and validity of the matching judgments using calibrated exemplars for the standardized assessment of writing, applied to performances from a broad age range across multiple genres and criteria.

**Reliability and validity.** The present findings provide further evidence that assessors are able to use calibrated exemplars to make quite reliable judgments at both the criterion and overall levels without extensive training in the procedure, and without any group moderation processes to establish consistency across assessors. Unlike the Stage 1 results, the level of reliability was somewhat moderated by genre type, criterion and grade level, although the

majority of reliability indices were at or above the common .8 correlation threshold for acceptable reliability in rubric scores (Brookhart & Chen, 2014).

The inter-judge correlations for the matching estimates were noticeably lower for the narrative performances, and particularly for the conventions criterion on these performances. This finding was corroborated by the lower correlations between each assessor's performance estimate and the average estimate of all other assessors for the 19 narrative performances with common marking. The reliability indices were also consistently lower than the levels of reliability observed for the calibrated exemplar method in Heldsinger and Humphry (2013) where it was applied to early-childhood writing. Similarly, the Stage 2 inter-judge correlations were noticeably lower for the grade 7 performances relative to the other grades. This finding is likely attributable, in part, to the lower observed variance for grade 7 estimates across the criteria relative to the other grades, which constrains the correlations. Further potential explanations for the reduced reliability of the matching judgments in the present study, and suggestions to enhance the reliability of the method are discussed in the Limitations and Future Research sections below.

A high level of concurrent validity was observed between the matching judgment estimates and the scores and sub-scores given by expert raters using the NAPLAN rubric. Again though, these correlations were somewhat lower for the narrative genre, the conventions criterion, and for grade 7, which may be attributable to the lower reliability of the matching estimates for these subsets of performances. Nevertheless, the majority of correlations between the matching estimates and rubric scores/ sub-scores exceeded .8, providing evidence that the criterion-level and overall matching estimates were valid and comparable to rubric scores. The individual grade level correlations were below the .8 criterion, which is likely due to the smaller variance and range of abilities within a grade level, as, for a given association, observed correlations are higher where the total variance is

greater. Moreover, this finding likely reflects the cognitive demands of matching each individual's performance against a relatively large number of exemplars, as discussed in the Limitations and Future Research sections below.

**Efficiency.** The median duration for all matching judgments in the entire task in Stage 2 was approximately six minutes. It was evident in the results that the assessors became, on average, substantially quicker at the matching judgments throughout the overall duration of the task. Allowing for an initial period during which assessors familiarized themselves with the calibrated exemplars, it was possible to attain an average judgment rate between 10 and 16 performances per hour across assessors. Moreover, the method alleviates some of the need for the extensive training and moderation that effective rubric scoring typically entails, although it is noted that all assessors involved in the matching task had separately undertaken NAPLAN rubric training and so they likely drew upon some of this experience when making their matching judgments, which may have enhanced the efficiency of these judgments (Heldsinger & Humphry, 2013; Jonsson & Svingby, 2007; Shermis, 2014). Nonetheless, a downward trend in median judgment duration across the matching exercise suggests that the increased efficiency was at least partly attributable to the assessors familiarizing themselves with the calibrated exemplars, and not based on their prior experience.

**Limitations.** Verbal feedback from assessors in Stage 2 clearly indicated that the matching process placed a relatively high cognitive load on them, particularly when more complex performances by older students were assessed. This issue was not as evident in the original application by Heldsinger and Humphry (2013) because they focused on early childhood writing with less sophisticated and complex features than the texts produced by older children in the present research. In addition, feedback and evaluation of the process by assessors indicated that too many calibrated exemplars were used in Stage 2. A total of 36 exemplars were selected so that the number of scores that could be assigned was similar to

the possible level of granularity from the NAPLAN analytic rubric. However, assessors did not always agree with the ordering of the calibrated exemplars that were presented, despite the very positive results from Stage 1 of the research.

In hindsight, this lack of consensus on the ordering of the exemplars was bound to be a limitation because the mean distance between the scale locations of adjacent exemplars was somewhat smaller than the mean standard error of the estimates on the scale established via comparative judgments. Consequently, in some cases there was insufficient certainty around whether adjacent exemplars were correctly ordered. Because the assessors were asked to match performances against calibrated exemplars, and they may not have perceived the exemplars to be ordered in a manner that was consistent with the overall solution from Stage 1, this discrepancy would have added to the difficulty of the decisions. This additional difficulty is likely to have had an adverse effect upon the reliability of the matching judgments.

In addition, one assessor annotated hard copies of performances to assist with the judgments and indicated to the researchers that brief descriptors of ranges of performance would have assisted the matching assessment process. As described in the Introduction, the research aimed to investigate whether calibrated exemplars could be used to capitalize on the advantages of comparative judgments while attaining greater efficiency using a two-stage process directly analogous to that adopted by Thurstone (1928). However, Thurstone's applications involved attitude and value statements, each of which were one sentence and could be rapidly read and ascertained. In contrast, because it is necessary to become familiar with a number of exemplars of extended writing and to hold information about these in mind while assessing a new performance, the use of calibrated exemplars to assess writing performances is more cognitively challenging. Therefore, any aids that can be provided to

assessors may make the process more efficient, as well as potentially increase the reliability and validity of the method.

The results indicated somewhat lower inter-judge correlations for narrative performances than persuasive performances. It is possible that the cognitive demands on the assessors were somewhat higher for narrative than persuasive performances. The persuasive texts generally have consistent macrostructural components because the instructions indicate students should include: (a) an introduction, (b) a body (with reasons in support); and (c) a conclusion. This greater consistency in macrostructure makes it a somewhat easier task for assessors to focus on key features of a persuasive performance when performing the matching. This macrostructural consistency may also explain the somewhat greater correlation between the authorial choices and conventions criteria in both the matching judgment estimates and rubric scores for the persuasive performances. In contrast, narrative texts have no standard macrostructural features, other than the use of paragraphs by sufficiently capable students. The assessment of narrative performances might also be more influenced by the subjective level of engagement of the assessor through creative elements such as ideas and voice (Grainger, Gooch, & Lambirth, 2005). For these reasons, there may be a greater need for descriptors and annotations for narrative than persuasive performances, but feedback from assessors indicated that such aids will be helpful for both types of writing.

Another issue raised by some assessors was that it seemed more difficult to make matching judgments for conventions, which may explain the generally lower levels of reliability and concurrent validity for this criterion. The reason indicated by these assessors was that it was possible to focus specifically on features such as spelling and sentence structure, as they are distinctive and separate, making it necessary to reconcile more than one distinctive judgment into an overall conventions judgment. In contrast, authorial choices comprise components that are more connected, such as ideas, audience and text cohesion. In

addition, it is possible that some of the components classified as conventions are more useful for distinguishing performances by the less capable students than the more capable students. The more capable students have generally mastered skills such as correct paragraphing and sentence structure, so assessors may have had somewhat greater clarity when assessing highly capable students on the authorial choices criterion than the conventions criterion, as differences in the former would be easier to discern in higher quality performances than differences in the latter.

**Future research.** Based on feedback and the current results, future research is planned in which: (a) fewer exemplars will be used, which are more definitively ordered for the assessors by taking explicit account of the standard error estimate for each of the exemplar performance estimates; and (b) theoretically and empirically informed descriptors will be provided in conjunction with exemplars to assessors to support matching judgments. The planned research will also provide the facility for annotations of calibrated exemplars to assist the assessors to recall key features and details of each performance, therefore reducing the cognitive load of the task. Finally, a larger sample that is representative of each grade, and for each genre at each grade in particular, will be recruited to obtain sufficiently robust reliability estimates within grades, as the current research was primarily focused upon obtaining reliability estimates across a range of grade levels.

These additions to the methodological design should make it possible to attain higher levels of inter-judge reliability with similar or improved efficiency to that found in Stage 2 of the present research. A drawback of these additions is that they require time and resources to implement and therefore somewhat detract from the greater efficiency of the method relative to rubric-based assessment. However, once the calibrated exemplars are scaffolded with descriptors and annotations, in particular, it is anticipated that new assessors will be able to



efficiently apply the method with an even higher level of reliability and validity than found in the present study in the absence of extensive training or moderation processes.

With respect to the inclusion of more specific judgment criteria in both the comparative judgment and matching stages of the research design, the present study went beyond the previous application by including two broad criterion for judgments based on feedback that aspects of these broad criteria were difficult to reconcile for certain writing performances. While the present research did not find evidence to motivate the inclusion of separate calibrated exemplar scales for each criterion, inclusion of more specific criteria in Stage 1 may motivate the inclusion of more specific calibrated exemplar scales for matching judgments.

The calibrated exemplar method is also applicable beyond writing assessments, as it is suitable for any sort of extended performance such as essays and media productions for which holistic judgments can be made (Heldsinger & Humphry, 2013). This broader applicability than either rubric-based or automated essay scoring is particularly pertinent given the current global trend toward more construct-relevant, performance-based assessments (Shermis, 2014). Consequently, future research is either planned or underway to demonstrate the robustness of the method for performance assessments more broadly. In particular, assessment tasks, including calibrated exemplars, have been developed to assess science investigations, informational reports, history, and mathematics. Software tools have also been concurrently developed to integrate the abovementioned descriptors and annotations into the matching process for these learning areas.

In addition, while the current findings demonstrate the potential application of the calibrated exemplar method to large-scale programs, particularly with the abovementioned improvements, this two-stage method was developed first and foremost to be used by classroom teachers and Heldsinger and Humphry (2013) provided evidence that teachers can

reliably and validly use the method with minimal training. Given the various criticisms of standardized assessments, including the narrowing of curricula, future research will examine whether the calibrated exemplar method can provide a valid and reliable basis for comparable performance assessments of classroom-based tasks across schools without the need for large-scale programs. This research requires providing a common bank of exemplars across classroom teachers and schools, which is already possible using the online software employed by the present study.

### **Summary and Conclusion**

To summarize, the aim of this research was to explore the reliability, validity and efficiency of the two-stage method of writing assessment in a standardized writing assessment context including multiple age groups, both narrative and persuasive genres, and two judgment criteria. The results showed good to very high levels of reliability in both stages of the method. The concurrent validity of the estimates from the second stage was high across criteria and genres, and the second stage was found to be reasonably efficient after an initial familiarization period. Nonetheless, the reliability and validity statistics were noticeably lower than a previous application of the two-stage method, particularly for the narrative performances on the conventions criterion.

These lower levels of reliability and concurrent validity are perhaps unsurprising given the more complex application in the present research. Future research will examine the reliability, validity and efficiency of the method with refinements to the method to help reduce the cognitive load of the task for assessors, and ultimately across a broader range of extended performances. Overall, the present research provides evidence for the viability of this method as a complement or alternative to existing methods of standardized writing assessment, and supports its potential for application in assessing extended performances across a range of learning areas.

## References

- Australian Curriculum, Assessment and Reporting Authority. (2010). *NAPLAN 2010 Writing Narrative Marking Guide*. Retrieved from the National Assessment Program website: [http://www.nap.edu.au/verve/\\_resources/2010\\_marking\\_guide.pdf](http://www.nap.edu.au/verve/_resources/2010_marking_guide.pdf)
- Australian Curriculum, Assessment and Reporting Authority. (2013). *NAPLAN 2013 Persuasive Writing Marking Guide*. Retrieved from the National Assessment Program website: [http://www.nap.edu.au/\\_resources/amended\\_2013\\_persuasive\\_writing\\_marking\\_guide\\_-with\\_cover.pdf](http://www.nap.edu.au/_resources/amended_2013_persuasive_writing_marking_guide_-with_cover.pdf)
- Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, 2(3), 451-462.
- Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills: Sage Publications, Inc.
- Bond, T. G., & Fox, C. M. (2012). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge.
- Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4), 324-345.
- Bramley, T., Bell, J. F., & Pollitt, A. (1998). Assessing Changes in Standards over Time Using Thurstone's Paired Comparisons. *Education Research and Perspectives*, 25(2), 1-24.
- Brindle, M., Graham, S., Harris, K. R., & Hebert, M. (2016). Third and fourth grade teacher's classroom practices in writing: A national survey. *Reading and Writing*, 29, 929-954.
- Brookhart, S. M., & Chen, F. (2014). The quality and effectiveness of descriptive rubrics. *Educational Review*, 1-26.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.

- Grainger, T., Gooouch, K., & Lambirth, A. (2005). *Creativity and writing: Developing voice and verve in the classroom*. Abingdon, UK: Routledge.
- Gyagenda, I. S., & Engelhard, G. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement*, 10(3), 225-246.
- Hayes, J. R., & Berninger, V. (2014). Cognitive processes in writing: A framework. In B. Arfé, J. Dockrell, & V. Berninger (Eds.), *Writing development and instruction in children with hearing loss, dyslexia, or oral language problems: Implications for assessment and instruction* (pp. 3-15). Oxford, UK: Oxford University Press.
- Heldsinger, S. A., & Humphry, S. M. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, 37(2), 1-19.
- Heldsinger, S. A., & Humphry, S. M. (2013). Using calibrated exemplars in the teacher-assessment of writing: an empirical study. *Educational Research*, 55(3), 219-235.
- Holme, B., & Humphry, S. M. (2008). *PairWise software*. Perth: University of Western Australia.
- House of Commons Education Committee. (2017). *Primary Assessment*. Retrieved from the Parliament UK website:  
<https://publications.parliament.uk/pa/cm201617/cmselect/cmeduc/682/682.pdf>
- Humphry, S. M., & Heldsinger, S. A. (2014). Common Structural Design Features of Rubrics May Represent a Threat to Validity. *Educational Researcher*, 43(5), 253-263.
- Humphry, S. M., & McGrane, J. A. (2015). Equating a large-scale writing assessment using pairwise comparisons of performances. *The Australian Educational Researcher*, 42(4), 443-460.

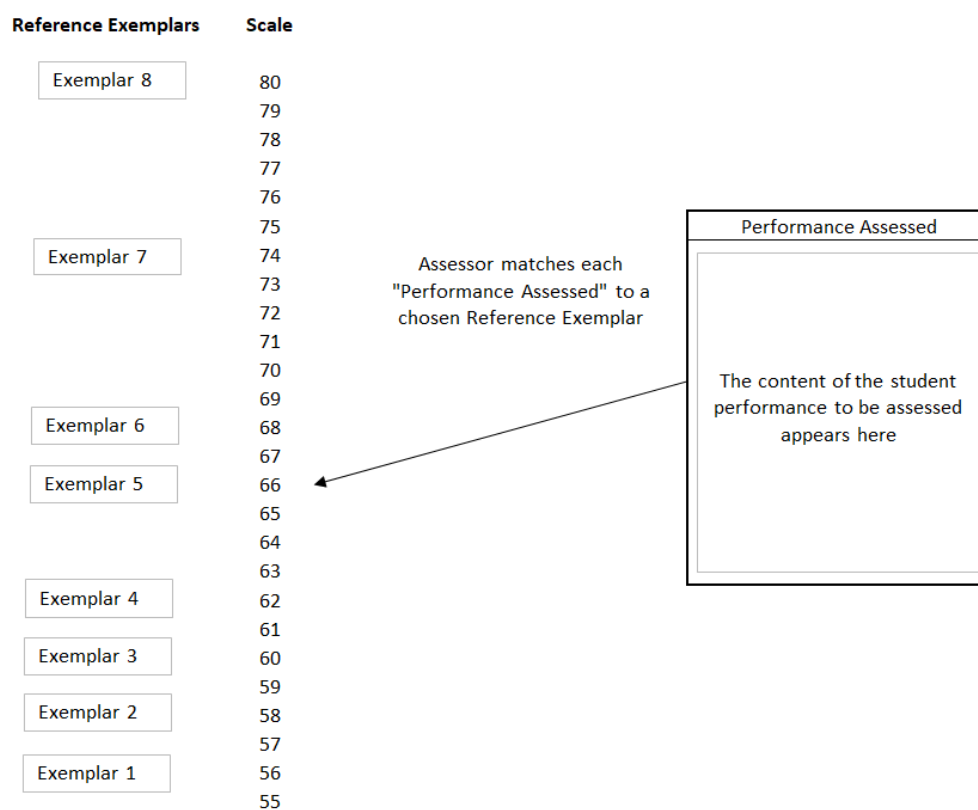
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, 2(2), 130-144.
- Luce, R. D. (1959). *Individual choice behavior: a theoretical analysis*. New York: Wiley.
- Mackenzie, N. M., Scull, J., & Munsie, L. (2013). Analysing writing: The development of a tool for use in the early years of schooling. *Issues in Educational Research*, 23(3), 375-393.
- McHale, K., & Cermak, S. A. (1992). Fine motor activities in elementary school: preliminary findings and provisional implications for children with fine motor problems. *American Journal of Occupational Therapy*, 46(10), 898-903.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: part I. *Journal of Applied Measurement*, 4(4), 386-422.
- Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3), 281-300.
- Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35(4), 435-448.
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18-39.
- Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, 34(2), 159-179.
- Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20, 53-76.
- Spandel, V. (2005). *Creating Writers Through 6-trait Writing: Assessment and Instruction* (4th ed.). Boston: Pearson and Allyn & Bacon.
- Steedle, J. T., & Ferrara, S. (2016). Evaluating Comparative Judgment as an Approach to Essay Scoring. *Applied Measurement in Education*, 29(3), 211-223.

- Tarricone, P., & Newhouse, C. P. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. *International Journal of Educational Technology in Higher Education*, 13(1), 1-16.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273-286.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529-554.
- Thurstone, L. L., & Chave, E. J. (1929). *The Measurement of Attitude: A Psychophysical Method and Some Experiments with a Scale for Measuring Attitudes Toward the Church*. Chicago, IL: University of Chicago Press.
- Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, 9(2), 1-10.
- Wyatt-Smith, C., & Jackson, C. (2016). NAPLAN data on writing: A picture of accelerating negative change. *Australian Journal of Language and Literacy*, 39(3), 233-244.

Table 1

*Summary of reliability and validity statistics across the two stages broken down by narrative (Nar.), persuasive (Pers.) and pooled (All) performances, as well as the combined, authorial choices and conventions criteria.*

	No. of performances			Statistic			Index
	All	Nar.	Pers.	All	Nar.	Pers.	
Stage 1: Reliability	160	80	80				
Combined				.982	.983	.980	Person Separation Index
Authorial				.971	.972	.968	
Conventions				.967	.971	.962	
Stage 2: Reliability	37	19	18				
Combined				.858	.779	.907	Mean inter-judge correlation
Authorial				.799	.710	.868	
Conventions				.781	.641	.852	
Stage 2: Validity	2380	1137	1243				
Combined				.872	.857	.900	Correlation with rubric score
Authorial				.851	.830	.882	
Conventions				.834	.819	.860	



*Figure 1.* Schematic of the digital display seen by assessors while making each of their matching judgments.