KEYNOTE

An innovative method for teachers to formatively assess writing online

Dr Sandy Heldsinger and Associate Professor Stephen Humphry
University of Western Australia

https://doi.org/10.37517/978-1-74286-685-7-1

Dr Sandy Heldsinger is leading the implemention of the Brightpath assessment and reporting software in schools. Sandy coordinated the Western Australia system-level assessments, has taught a masters-level course in educational assessment for a number of years and has led the development of a wide range of resources, including reporting software, to support schools in using assessment to improve student performance. Sandy recently developed central components of the Western Australian Curriculum Outline for WA's School Curriculum and Standards Authority and she coauthored What Teachers Need To Know About Assessment and Reporting*, published by ACER.*

Steve Humphry is an Associate Professor in Educational Assessment, Measurement and Evaluation at The University of Western Australia (UWA). He won a position with UWA in 2006 and has held substantial Australian Research Council Linkage grants continuously since 2008. He has worked with industry partners on these grants, including the Australian Curriculum Assessment and Reporting Authority, UNESCO's International Institute for Educational Planning, and the School Curriculum and Standards Authority. Stephen has extensive experience in the NAPLAN and WALNA (Western Australian Literacy and Numeracy Assessment) large-scale testing programs. His research has focused increasingly on developing novel approaches that allow classroom teachers to reliably assess students in areas not amenable to large-scale testing such as visual art and science investigations. In industry work, between 2011 and 2013, Stephen led the Central Analysis of Data project for NAPLAN, on which a team of psychometricians established NAPLAN literacy and numeracy scales that form the basis for reporting on performance at the Australian, state and territory, school and individual student levels.

Abstract

Assessment is an integral component of effective teaching and a teacher's professional judgement influences all routine aspects of their work. In the last 20 years, there has been considerable work internationally to support teachers in using assessment to improve student learning. However, there is a pressing issue that impedes teachers' professional judgement being exploited to its full potential. The issue relates to teacher assessment of learning progression in the context of extended performances such as essays and arises from the complexity of obtaining reliable or consistent teacher assessments of students' work. Literature published in the United States, England and Australia details evidence of low reliability and bias in teacher assessments. As a result, despite policymakers' willingness to consider making greater use of teachers' judgements in summative assessment, and thus provide for greater parity of esteem between teacher assessments and standardised testing, few gains have been made. Although low reliability of scoring is a pressing issue in contexts where the data are used for summative purposes, it is also an issue for formative assessment. Inaccurate assessment necessarily impedes the effectiveness of any follow-up activity, and hence the effectiveness of formative assessment. In this session, we share our research of writing assessment and explain how it has led to the development of an innovative assessment process that provides the advantages of rubrics, comparative judgements, and automated marking with few of the disadvantages.

Introduction

Despite the widespread desire for teacher judgements to be used for summative assessments, attaining reliable judgements has been a challenge (Brookhart, 2013; Harlen, 2004; Johnson, 2013). Instead, external standardised assessments are mostly used for this purpose. Similarly, although assessment is an integral component of teaching, and professional judgement influences various aspects of teachers' work (Black & Wiliam, 2010; Du Four, 2007; Hattie & Timperley, 2007), the reliability of formative assessments is seldom examined.

A growing body of research shows that teachers make reliable judgements by making pairwise comparisons of extended performances (Humphry & Heldsinger, 2019). Using this approach, teachers compare pairs of performances and judge which performance, in each pair, demonstrates a higher level of attainment. Performances are placed on the scale from weakest to strongest, empirically showing learning progression. The terms comparative judgement, comparative pairs, and paired comparison are also used to describe pairwise comparisons (Tarricone & Newhouse, 2016).

A drawback of pairwise comparisons is that they are time-consuming as a method for teacher assessment (Bramley et al., 1998). In addition, pairwise comparisons provide the basis for scaling and ordering of student performances but the approach does not directly avail teachers of diagnostic information in a form that can be acted upon. However, once performances have been ordered, they can be qualitatively examined and doing so provides insight into changes in features of writing observed with increasing development. Thus, the application of pairwise comparisons potentially provides the basis both for internally consistent judgements and diagnostic information. The method described in this article is designed to provide these advantages to classroom teachers. As described to follow, the two-staged method is designed so that it is time-effective, accessible, and provides immediate and actionable formative feedback.

The two-stage assessment method

Constructing scales using pairwise comparisons

The first stage in the two-stage assessment method is to calibrate a scale and subsequently to select exemplars. The literature provides background on the use of the method of pairwise comparisons in education and other fields (Bond & Fox, 2001; Bramley et al., 1998; Pollitt, 2012; Thurstone, 1927, 1959). In Stage 1, assessment tasks are administered by classroom teachers and a relatively large number of performances are collected. Teachers compare these performances and select which performance is on-balance better given key performance features to be considered. The pairwise comparison data are analysed using the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959) to produce a performance scale. Scale locations are inferred from the proportions of judgements in favour of each performance when compared with others. If all performances were compared with each other, the strongest performance would be the one judged better than the other performances the greatest number of times.

However, in practice, scaling techniques can be used and it is unnecessary for each performance to be compared with every other performance. Data from pairwise comparisons are analysed to examine the overall internal consistency. Data are also analysed to ascertain whether each teacher's comparisons are consistent with the overall scale locations, within expectations given the BTL model. Specifically, the Person Separation Index is used to examine internal consistency and fit indices are used to examine the consistency of teachers' comparisons with overall scale locations, as reported, for example, in Humphry and Heldsinger (2020).

Fit to the BTL model is also examined for each performance on the scale. Fit indices and qualitative examination of performances are used to select a set of exemplars for use in Stage 2, in which teachers assess other performances against the scale. Performances are not used as exemplars if the pairwise comparisons produce data that departs too much from the Guttman pattern (Guttman, 1944). Performances with relatively Guttman-like patterns are compared consistently with overall ordering and provide better exemplars for Stage 2. These performances are more clearly ordered and provide a clearer reference point for assessment against the scale.

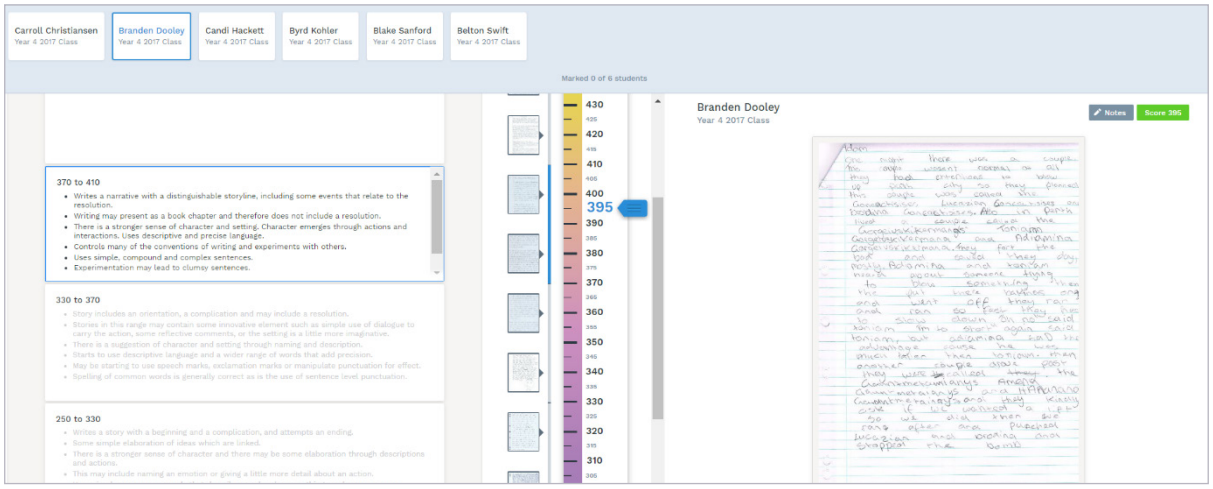
Exemplars and descriptors

Once performances have been placed on a scale, they are placed in order of location on the scale physically (e.g. on a table) and examined to infer the features and levels of writing that students in a given range of the scale demonstrate in their performances. In this way, learning progressions are described. In contrast with typical rubrics, performance descriptors are based on a systematic analysis of the performances placed in order according to one or more criteria. This enables teachers to glean empirically based information about how performances change with progression from lower to higher levels. It also provides a basis for specific feedback on key points, which are referred to as Teaching Points. Descriptors focus on features that are most relevant in a given range of the scale.

The teacher's ruler is an interactive display comprising several key elements, as shown in Figure 1. Teachers assess their own students' performances, shown on the right-hand-side, against the ordered exemplars, which are shown in the centre of the screen. Teachers refer to the empirically based descriptors displayed on the left-hand side. Teachers can click on exemplars to expand and view them on the left-hand side of the display. To assess their students' work, teachers locate where a performance is likely to sit on the scale based on comparisons with the exemplars and using the descriptors as a guide.

The interactive display provides the advantages of rubrics and comparative judgements. Specifically, teachers refer to general descriptions of performances relating to a given range, and they also compare performances with real, pre-calibrated exemplars.

Figure 1 Teacher's Ruler display



Teachers make an on-balance judgement based on their analysis of the strengths and weaknesses of the performance, and to determine which exemplar the performance was closest to or which two exemplars it fell between. In the display shown in Figure 1, exemplars are displayed in order from lowest to highest. Teacher comparisons are implicit rather than explicit. For example, if a performance is judged above the 10th exemplar but below the 11th exemplar in the Teacher's Ruler, it is implied that the performance is better than all exemplars below the 10th and worse than all exemplars above the 11th. The performance is given the scale score associated with or above or below the exemplar. The scale is shown in the centre of Figure 1. The scale locations are based on the analysis of the pairwise comparison data; specifically, they are transformations of the logit scale obtained from analysis of data using the BTL model.

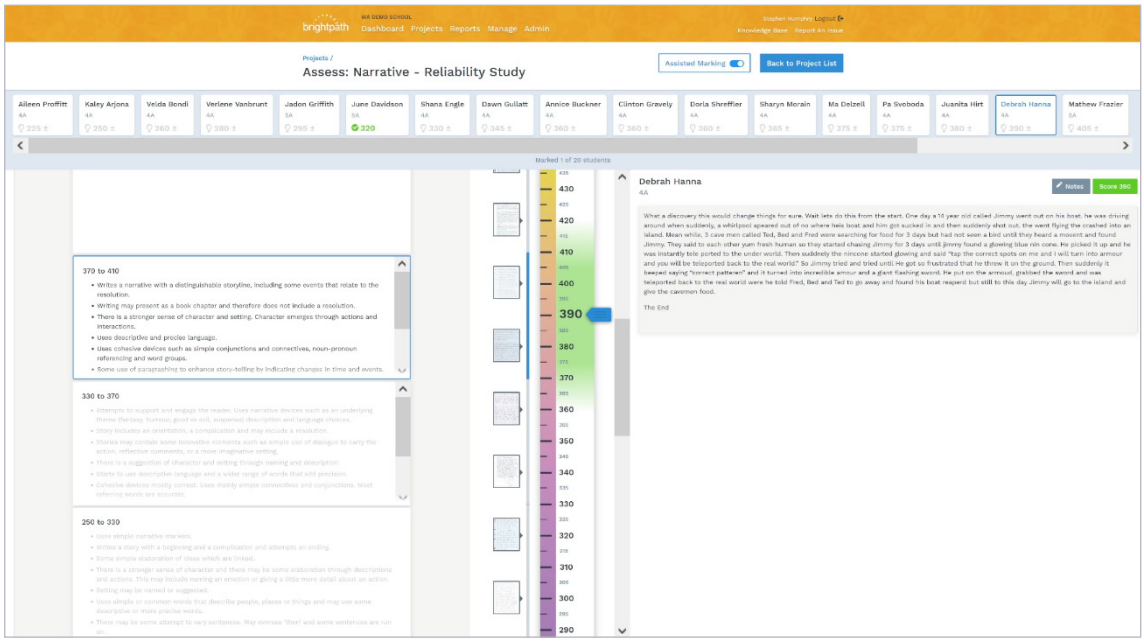
Judges are provided with a guide to help make their judgements. This guide contains all the calibrated exemplars, the performance descriptors, and a close qualitative analysis of each exemplar.

Assisted marking with Natural Language Processing and calibrated exemplars

Automated scoring is often used instead of human marking or to check human marking. However, it is not necessary to adopt a process in which human and automated scoring occur separately. An automated Marking Assistant has been designed to help teachers quickly focus in the right zone of the Teacher's Ruler in much the same way that a search-suggestion helps users focus on information that is most relevant. This process is designed to help teachers to concentrate on features of writing that are best judged by humans.

Based on Natural Language Processing (NLP) indices, the automated assistant predicts scale locations of online performances as a starting point. In Figure 2, the blue arrow pointer shows the predicted score for the performance on the right. The green zone shows the interval or range in which the performance's score is predicted to lie with approximately 70 per cent confidence. For teachers who are not yet familiar with the exemplars, the predictions enable efficient assessments from the start, while teachers gain familiarity. For teachers who are familiar, the predictions serve as a reference point. Teachers can also turn the predictions off.

Figure 2 Marking Assistant prediction on the display



Moderation

The Teacher's Ruler uses exemplars to show what constitutes a given score. Teachers in one school or classroom in a school see the same set of ordered exemplars as teachers in another school or classroom. If the assessments are conducted effectively, they are automatically moderated. Various issues may occur with rubrics depending on how they are designed and constructed (Humphry & Heldsinger, 2014). Bias and rating tendencies that are common in rubrics are limited by having exemplars as the basis for implicit pairwise comparisons. In rubrics, performances are referenced to descriptions, and the descriptions can be interpreted differently by different teachers. In pairwise comparisons, it is difficult to introduce bias because one performance is directly compared with another.

Formative assessment

Consistent with the logic of the cumulative ordering in the Guttman pattern (Guttman, 1944), the following approach is used. For students in a given range of the Teacher's Ruler scale, descriptors applicable to students in the next higher range are used as teaching points. The rationale is that descriptors in the next higher range are most likely to describe what lies in a student's zone of proximal development with learning progression.

Teachers can refer to specific features of exemplars in providing feedback to students. Descriptors convey a general sense of performance; whereas the exemplars show, in more tangible and specific terms, what performances at a given level look like. The exemplars explicitly show different levels of performances in a way that is difficult to fully capture using descriptions of the kind that appear in rubrics. Together, descriptors and exemplars convey learning progression better than either individually.

Reliability of teacher judgements

Several studies have been conducted to examine the reliability of teacher judgements of narrative, persuasive, and information-report writing assessment using the second stage of the two-stage method. In each of these studies, all participants assessed a common set of approximately 25 performances using the Teacher's Ruler. Each marker's scale scores for the common performances were compared with the average scale scores given by all the other markers in the study. The correlations obtained from these studies are shown in Table 1 and show high levels of reliability using the Teacher's Ruler to assess the extended performances.

Table 1 Summary of reliability of the second stage assessments in a number of studies

	Study 1	Study 2	Other studies
Narrative	0.903*	0.927*	0.938
	n=12	n=37	n=65
Persuasive	0.848*	0.925	
	n=8	n=30	
Information report	0.966		
	n=34		

*previously published results

The evidence reported here is collected without use of Assisted Marking. Some of these results have been reported in published literature (Heldsinger & Humphry, 2013; Humphry et al., 2017; Humphry & Heldsinger, 2019; Humphry & Heldsinger, 2020).

Discussion

> One requirement that has been present since the very first version, over a half century ago, is that tests should be adequately documented, the procedures by which tests were developed should be documented, and evidence regarding the validity of the tests, and specifically the reliability, must be produced (Black & Wiliam, 2012, p. 252).

Discussion of reliability in the context of teachers' assessments is often referred to as inter-rater reliability and relates to the generalisability of scores across markers or scorers. Differences that arise in scores that are not a function of student ability, but from differences in examiners, constitute a source of measurement error that negatively impact the reliability of the assessment. The key implication is that where differences in scores more accurately represent the differences in the construct being assessed, people can have more trust in scores when drawing inferences about students' ability and making decisions about follow-up actions.

The results outlined in this article provide empirical evidence that the two-stage method enables reliable teacher assessments, responding to calls for research into the reliability of teacher assessments by Harlen (2005), Brookhart (2013), and Johnson (2013). A negative impact of giving high-profile to external and standardised test-based results can be a loss of assessment skill on the part of teachers as well as loss of confidence in their ability to make sound assessments of their students (Black et al., 2010, 2011). A significant reason for placing emphasis on external and standardised assessments is the belief that teacher assessments are not sufficiently reliable. The key benefit of enabling teachers to make reliable assessments is that the professionalism of teachers is valued and fostered, consistent with the general desire to value teacher judgements observed by Johnson (2013).

In conclusion, the two-stage method of assessment enables teachers to make reliable judgements of writing. An advantage of the method over pairwise comparisons alone is that once a scale has been constructed, the average time to assess a performance is reasonably modest. The use of assisted marking further reduces assessment time by enabling teachers to focus on what they are best placed to assess in performances. Unlike external testing programs, by using the two-stage method classroom teachers assess their own students and provide formative feedback based on their own assessments and familiarity with the students' work.

References

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice*, *17*, 217-234. https://doi.org/10.1080/09695941003696016

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice*, *18*, 451–469. https://doi.org/10.1080/096959 4X.2011.557020

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90. https://doi.org/10.1177/003172171009200119

Black, P., & Wiliam, D. (2012). The reliability of assessments. In J. Gardner (Ed.), *Assessment and learning* (2nd Edn, pp. 243–263). Sage Publications Ltd.

Bond, T. G., & Fox, C. M. (Eds.) (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates Inc. https://doi.org/10.4324/9781410600127

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.

Brookhart, S. M. (2013). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 69–90. https://doi.org/10.1080/0969594X.2012.703170

Du Four, R. (2007). Once upon a time: A tale of excellence in assessment. In D. Reeves (Ed.), *Ahead of the curve. The power of assessment to transform teaching and learning* (pp. 253–267). Solution Tree Press.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*(2), 139–150.

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

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*(3), 245–270. https://doi.org/10.1080/02671520500193744

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heldsinger, S., & Humphry, S. (2013). Using calibrated exemplars in the teacher-assessment of writing: an empirical study. *Educational Research, 55*(3), 219–235. https://doi.org/10.1080/00131881.2013.825159

Humphry, S., & Heldsinger, S. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher*, *43*(5), 253–263.

Humphry, S., & Heldsinger, S. (2019). A two-stage method for classroom assessments of essay writing. *Journal of Educational Measurement*, *56*(3), 505–520. https://doi.org/10.1111/jedm.12223

Humphry, S., & Heldsinger, S. (2020). A two-stage method for obtaining reliable teacher assessments of writing. *Frontiers in Education*, *5*. https://doi.org/10.3389/feduc.2020.00006

Humphry, S., Heldsinger, S., & Dawkins, S. (2017). A two-stage assessment method for assessing oral language in early childhood. *Australian Journal of Education*, *61*(2), 124–140. https://doi.org10.1177/0004944117712777

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education*, *28*(1), 91–105. https://doi.org/10.1080/02671522.2012.754229

Luce, R. D. (1959). *Individual choice behaviours: A theoretical analysis*. J.Wiley.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

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), 16. https://doi.org/10.1186/s41239-016-0018-x

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286. https://doi.org/10.1037/h0070288

Thurstone, L. L. (1959). *The measurement of values*. The University of Chicago Press.