The classification accuracy and consistency of comparative judgement of writing compared to rubric-based teacher assessment

Christopher Wheadon^a, Anne Pinot de Moira^b, Daisy Christodoulou^a

^a *No More Marking, Durham, UK;* ^b *Independent Consultant, Surrey, UK*

Summary

Writing is generally assessed internationally using rubric-based approaches, but there is a growing body of evidence to suggest that the reliability of such approaches is poor. In contrast, comparative judgement studies suggest that it is possible to assess open ended tasks such as writing with greater reliability. Many previous studies, however, have failed to provide direct comparisons between these approaches as the reliability measures for rubric- and marking-based studies are not comparable with the internal measures of reliability cited by comparative judgement studies. We investigated the classification accuracy and consistency of a rubric-based approach to the grading of writing with a comparative judgement approach. The writing was gathered from 11-year-olds in low stakes settings in England. We present evidence that the comparative judgement approach has twice the classification accuracy as the rubric-based approach and is perfectly viable in terms of its efficiency. We discuss the limitations of the comparisons and consider what a national system for assessing writing based on a comparative judgement approach could look like.

Keywords: Comparative judgement; Assessment; Writing assessment; Classification accuracy; Classification consistency

^a Corresponding author: Christopher Wheadon, Founder & CEO No More Marking Ltd, Durham, UK, chris@nomoremarking.com

Research aims

The main aims of this study are to compare the classification accuracy and consistency of two approaches to the grading of writing and to consider the extent to which these approaches measure the same construct. The first approach is traditional rubric-based grading which, in England, is operationalised at primary education level through the Teacher Assessment Framework (TAF) (Cuff et al., 2018). Using the TAF, teachers in England grade students at the end of their primary education (aged 11, Year 6) with three grades according to the rubric (Greater Depth Standard, Expected Standard, Working Towards Standard). The teachers consider a range of writing produced by their students and decide on a grade. In any one year, 25% of schools are visited by moderators who view a minimum of 15% of the cohort, and the grades of these students might be revised. The moderation process is described as a professional discussion through which agreement is reached (Standards & Testing Agency, 2019). The writing assessment is 'high-stakes' as outcomes form part of primary school accountability measures but there is no research evidence on the classification accuracy and consistency of this approach.

The second approach to grading is the use of comparative judgement (CJ). The CJ approach is operationalised in England using a commercial programme called *Assessing Primary Writing* (Wheadon et al., 2020). Schools pay to take part in the project, which allows them to gain nationally standardised measures of writing ability for their students across the primary phase of education. All students sit the same writing tasks in strictly controlled conditions, completing their writing within one hour. Teachers from the schools then take part in judging the writing. To ensure that the judging is not biased, and as part of an equating design, every fifth judgement made by a teacher compares writing from two other schools. Teachers never compare their own students with students from another school; thus the scores the students receive are standardised and fair (Wheadon et al., 2020).

There are clear differences between the two approaches to assessing writing in what is considered by the teachers. Teachers using the TAF have very clear specifications of what to look for and they look for it in portfolios of writing that may have been substantially supported and revised in high-stakes conditions. Teachers using CJ are given a very open question: which script has the better writing? They judge single snapshots of writing that have taken place in controlled but low-stakes conditions. Despite these differences in mechanics, however, it is worth attempting to discover whether the underlying construct that both seek to measure is the same and compare how well each method can replicate the grades that are produced. Arguments for or against the relative validity of one or the approach in measuring what we value in writing would be undermined if the reliability was not sufficient to support its use.

Background

Rubric-based approaches to the marking and grading of writing are the norm internationally but there is some evidence that the marking and grading consistency of this approach is poor. In England, the reliability of a now defunct writing test for students aged 11 was estimated to have a +/- 5 mark accuracy on a mark range of 43 (He et al., 2013). Research on the marking of writing of older students has found similarly poor marking consistency, with levels comparable in England to internationally (Rhead et al., 2018). While there is almost no evidence regarding the consistency of

direct classification of writing into grades rather than a fine mark, there is no theoretical reason to believe that a broad classification would be more reliable than a fine mark that is then converted into a grade (Wheadon & Pinot de Moira, 2013). Attempts to improve mark schemes and rubrics have led to no reported gains in marking consistency, rather issues arise with hyper-specificity distorting approaches to teaching and learning (Cuff et al., 2018; Humphry & Heldsinger, 2014; Rezaei & Lovorn, 2010).

In contrast to rubric-based approaches, there is a growing body of evidence to support the use of comparative judgement as a reliable alternative for the assessment of open-ended tasks. With CJ, judges compare two pieces of work side by side and decide on the better piece according to the criteria they have been set. Following repeated comparisons, the Bradley Terry model (Bradley & Terry, 1952; Luce, 1959) is used to model scores that represent the underlying quality of scripts. The scores can be equated and scaled as they share the essential properties of the Rasch model (Andrich, 1978).

The reliability measures cited in CJ studies has generally been the scale separation reliability (SSR) coefficient, defined in the same way as the person separation reliability in Rasch modelling (Andrich, 1982). The SSR values cited in the literature on CJ are high: 0.98 for student writing performances in the NAPLAN Writing tests for grades 3, 5, 7 and 9 in Australia (Humphry & McGrane, 2015); 0.98 for primary school narratives (Heldsinger & Humphry, 2010); 0.84 for an essay assessing skills in academic writing (van Daal et al., 2019). Unfortunately, it is often hard to compare the SSR of CJ approaches with more traditional rubric-based approaches as the SSR coefficient is not directly comparable to the inter-rater reliability coefficient reported by traditional marking studies (Bramley & Vitello, 2019). Two studies, however, have directly investigated how well traditional marking compares to CJ in predicting results in a separate assessment. In both studies, the research found the predictive value of CJ exercises to be better than single marking, and similar to that of double marking (Benton & Gallacher, 2018; Holmes et al., 2017). The original intention of this study was to predict results in a reading test from the two approaches to the grading of writing, however, due to COVID-19 all the national 2020 reading tests were cancelled restricting the scope of this study. Instead, classification accuracy and consistency measures are used to provide direct comparisons between the CJ approach and the rubric based approach.

Classification accuracy and consistency

As described above, one of the key issues with comparing reliability results from traditional approaches to marking or grading and CJ approaches has been the different metrics used for each. As the writing in this study is graded, however, classification accuracy and consistency measures of reliability can be used to make such direct comparisons.

Classification accuracy conceptualises the chance a script will be given a grade that reflects its true worth under repeated administrations of an assessment (Wheadon, 2014). It is often evaluated by false positive and false negative error rates (Hanson & Brennan, 1990; Lee et al., 2002). The conditional false positive error rate is defined as the probability that a student is classified into a grade that is higher than his or her true grade. Conversely, the conditional false negative error rate is the probability that a student is classified into a grade that is lower than his or her true grade.

Classification accuracy is then determined by subtracting the incidence of false positives and false negatives from one.

A student's true grade is unmeasurable but, nevertheless, most schemes of assessment can give rise to a measure that might be termed the definitive grade. The structure for national moderation using the TAF is such that the definitive grade for a script is that which would be determined by the senior moderator. Repeated gradings of the same scripts then allow estimations of conditional false positive and false negative error rates.

Under Item Response Theory models, the definitive grade is defined as that given by the initial model fitted to the data (Wheadon, 2014). The probability of success on each item is algebraically equivalent to the probability of a judge selecting one or other scripts in a CJ decision (Andrich, 1978), so Equation 1 can be used to estimate the expected score distribution for scripts in a CJ study. Multiple iterations over this score distributions result in multiple grade distributions which allow estimations of conditional false positive and false negative error rates.

$$P(i > j) = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}}$$

Equation 1

The classification consistency of a test is a measure of the probability that a script will be given the same grade under repeated administrations of an assessment (Wheadon, 2014). While many measures of classification consistency are available, Krippendorff's alpha (K- α) is most appropriate for the data from the study (Hayes & Krippendorff, 2007). The measure is reported on the scale zero to one: no consistency to perfect consistency. It copes with missing data; an important factor because in this study not all scripts were seen by all participants. It also adjusts for chance agreement, is invariant to the number of participants and is appropriate for nominal data such as grades.

Study design

Data

Nine schools were recruited to share their Year 6 writing submitted as part of the March 2020 *Assessing Primary Writing* window. The schools were incentivised with a free *No More Marking* subscription in the 2020/2021 academic year. In return, they committed to supplying additional information about their students. This included gender, Pupil Premium status¹ and English as an Additional Language (EAL) status. They were also asked to rank the writing ability of their students before the March 2020 *Assessing Primary Writing* window and to predict a final Key Stage 2 (KS2) writing grade.

The schools were sampled from the pool of those new to *Assessing Primary Writing* in the 2019/2020 academic year. This was in order that both the writing and the teacher KS2 prediction were unaffected by previous preparation for CJ tasks. Between them the schools provided scripts from 349 students. The scripts were used as the basis for comparison between CJ and TAF grading.

¹ Pupil premium is a grant designed to allow schools to help disadvantaged pupils by improving their progress and the exam results they achieve.

The schools were selected using a cluster sample to represent a range of intakes. Independent and special schools were excluded as were schools with low student numbers. Not all the schools initially sampled agreed to take part in the study but replacements were selected randomly from within the original clusters. One school dropped out late on in the study and therefore the final sample, described in Table 1, shows slightly fewer students from schools in the lower performing cluster.

Table 1 Sample of schools included in the study

Cluster	Region	School Type	Year 6 students (Scripts)
High performing	Yorkshire	Foundation school	46
High performing	Yorkshire	Academy converter	28
High performing	North West	Academy sponsor led	26
Mid performing	Yorkshire	Academy converter	28
Mid performing	East of England	Foundation school	57
Mid performing	London	Community school	59
Mid performing	North West	Voluntary controlled school	39
Low performing	North West	Academy sponsor led	13
Low performing	Yorkshire	Academy converter	53

Grading using the teacher assessment framework

Participants

The scripts used in *Assessing Primary Writing* are not normally graded using the TAF and therefore this aspect of the study required additional data collection. To this end, 30 teachers were recruited to complete a grading exercise.

To allow effective checks on teacher credentials, the invitation to take part in the study was only emailed to schools that currently subscribe to a *No More Marking* product. Potential participants were required to have recent experience in using the TAF to grade student writing. There was considerable interest in the study and within 24 hours over 300 teachers had responded. The pool was restricted to those working in England on the assumption that these teachers would have the greatest experience of the KS2 curriculum. Thirty teachers were randomly selected from those that remained. The sample included teachers from across the whole country, teaching in a wide range of school types with very different KS2 results.

The participants were paid for their role and in return were required to grade a small number of pre-standardisation scripts, to attend a standardisation webinar and, afterwards, to grade 117 scripts using an online tool.

Senior moderator

A senior moderator was recruited to prepare and present the standardisation webinar and to oversee the grading of scripts. This moderator had experience as a Local Authority Lead Moderator and as an Operational External Moderator and Deputy Standardisation Team Leader for the Standards and Testing Agency. Before attending the webinar, teachers had to work through a set of training materials that are normally provided as part of the national moderation training for moderators. They also had to grade nine scripts in a pre-standardisation exercise. The teachers were

given feedback on their grading. The webinar talked through details of the TAF. It discussed how to implement the TAF with reference to specific pieces of writing chosen from the national training materials and, in addition, from the recent *Assessing Primary Writing* window. Participants could ask questions and clarify areas of uncertainty.

As well as training the recruited participants in the grading exercise, the senior moderator graded all 349 scripts included in the study.

Grading

Each participant was given a unique pack of scripts to ensure that, as far as possible, each script was graded the same number of times. No script was viewed on fewer than eight occasions nor on more than 11. The median number of views was ten.

All the grading was completed using an online platform. The participants were presented with scanned scripts and asked to assign a grade. The TAF normally requires that a student should be categorised as *Working Towards the Standard* (WTS), at *Expected Standard* (EXS) or working at *Greater Depth Standard* (GDS). In the study, we asked for a more granular judgement. Each of the three TAF categories was further subdivided into three, a, b or c, to denote high, mid and low performances. There were therefore nine fine grades that could be assigned to a script, GDSa, GDSb, GDSc, EXSa, EXSb, EXSc, WTSa, WTSb and WTSc. For the study, the normal TAF grades are referred to as KS2 grades while the nine grades are referred to as fine grades.

Comparative judgements

Comparative judgements of all scripts included in the study were made as part of the normal *Assessing Primary Writing* process by teachers from each subscribed school. In the study, each script was involved in a local comparison an average of 16 times. Nearly a quarter of the scripts were also involved in moderation comparisons.

The CJ process produces a scaled score for every script and this score is standardised across all schools. Therefore, standardised CJ scores reported in this study are comparable between all the study schools. They are reported on an arbitrary scale ranging between 300 and 700. For the purposes of this study, standardised CJ scores were converted to grades by equipercentile equating to the grade distribution generated from the median TAF grade awarded to the scripts.

Analysis

Descriptive data

The timing of the Year 6 *Assessing Primary Writing* window is such that work is submitted about three months before the teacher assessment of writing. Furthermore, the CJ are based on a single piece of evidence rather than a portfolio. Therefore, it is unsurprising that the TAF grades awarded to the study scripts were lower than might be seen in the final assessment. Nationally in 2019, 20% of students were awarded GDS and 78% were awarded EXS or higher. Table 2 shows the median TAF grade distribution. Equipercentile points from this distribution were used to convert the CJ scores to grades. While the median grade for a script was never greater than GDSc, several scripts were awarded higher grades by individual moderators.

Table 2 The median TAF grade distribution for the scripts included in the study

Grade	GDSa	GDSb	GDSc	EXSa	EXSb	EXSc	WTSa	WTSb	WTSc
Cumulative %	0.00	0.00	1.72	5.73	23.50	50.43	73.93	91.98	100.00

For each school, Figure 1 shows the relationship between the CJ scaled score for a script and the rank assigned to a candidate by the class teacher. There is a clear positive relationship between the two measures. The KS2 grade is overlaid in colour; a light blue circle denoting WTS, a dark blue triangle, EXS and a green square, GDS. Note that the TAF KS2 grade is randomly chosen from all the grades awarded to a given script as part of the study design. Over 70% of scripts were awarded the same KS2 grades by CJ and the TAF and, with a Kendall's τ_b correlation coefficient of 0.505 between the two grading approaches, there is some evidence they are measuring the same construct. Indeed, the correlation between the median TAF grade and the CJ grade was higher still at 0.639, with nearly 80% of grades awarded being in agreement.

Figure 1 Comparative judgement scaled scores plotted against teacher rank, cross-classified by single moderator TAF grade

Table 3 shows the mean KS2 grade for the two different grading approaches and for the teacher predictions. It is subdivided by the different demographic characteristics of students included in the study. The teacher predictions are higher than those from the study because they reflect the expected level at the end of Year 6 based on a portfolio of work. A set of two-way analyses of variance reveal no evidence of any interaction effect between gender, pupil premium eligibility or EAL and the grading approach. In other words, neither CJ, TAF nor teacher grades appear to exhibit bias with respect to the different demographic characteristics.

Table 3 Mean KS2 grade awarded dependent on grading approach and demographic characteristics

	CJ		TAF		Teacher prediction		Students
	Mean	se	Mean	se	Mean	se	
Female	1.61	0.04	1.63	0.01	2.15	0.05	172
Male	1.44	0.04	1.49	0.01	1.94	0.05	177
Not PP	1.60	0.04	1.63	0.01	2.12	0.05	217
PP	1.39	0.04	1.44	0.01	1.91	0.05	132
Not EAL	1.50	0.03	1.56	0.01	2.05	0.04	270
EAL	1.59	0.06	1.56	0.02	2.01	0.08	79

Speed of grading

For any system of measurement to be credible, it must be demonstrably valid but also administratively practical. Most assessment systems devote considerable resources to the delivery of outcome measures. Moderating, examining or judging scripts requires an investment of time which adds delay and cost to the system.

When scripts are graded using the TAF, they need only to be viewed by one moderator before an outcome measure can be produced. On the other hand, for CJ, the scripts must be viewed by multiple judges.

In this study, there was no exact measure of the time to grade each script using the TAF. However, the senior moderator, who viewed the complete sample, estimated that each script took approximately two minutes to grade. The average essay length for the Year 6 *Assessing Primary Writing* task used in this study was about 240 words² and so, crudely, the moderators were processing 120 words per minute. In a similar study using Year 11 writing, which is inherently more complex, Benton & Gallacher (2018) suggested a processing speed of about 70 words per minute. Given the difference in the Year 6 and Year 11 writing tasks, therefore, the senior moderator's estimate of two minutes per script appears plausible.

In contrast, there were exact measurements of the time taken for each judgement made using CJ. These measurements were combined to estimate the total time devoted to judging a single script. For the sample included in the study, the median time invested per script was four minutes. In other words, grading a single script using CJ proved equivalent, in terms of time taken, to having two moderators grade the same script using the TAF.

² Based on a random 5% sample of the study scripts.

That CJ proves more time consuming than applying a single grade to a script is not in itself news. Benton & Gallacher (2018), presented starker findings suggesting that the time taken to grade a script using CJ was roughly equivalent to triple marking the same script. A similar study by Ofqual concluded that, for equivalent levels of predictive accuracy, the workload for CJ was somewhere between that of double and triple marking (Holmes et al., 2017). Of interest therefore is how, in the context of KS2 assessment, the extra investment of time affects the classification accuracy and consistency of grade outcomes.

Classification accuracy

Table 4 shows classification accuracy expressed as a percentage for single moderator TAF grades and for grades derived from CJ. The CJ provides greater agreement with the definitive grade. For the fine grades, the probability of exact agreement between grades using CJ is almost double that using the TAF. The result, while striking, is not surprising. It is consistent with the literature on CJ, but also theoretically predictable. Any single moderator is likely to bring personal bias to their grading, which is not shared by other moderators. This bias will reduce the agreement between moderators. CJ uses a more consensual approach in which individual biases will cancel out, assuming a large enough pool of judges. Two sets of judges, where the individual bias has been cancelled, out are therefore more likely to agree with each other.

Table 4 Classification accuracy for CJ and TAF grading (expressed as a percentage)

	Fine grades			KS2 grades		
	False positive	Exact agreement	False negative	False positive	Exact agreement	False negative
TAF	30.8	33.0	36.2	16.6	70.2	13.2
CJ	18.5	62.6	18.9	5.9	88.2	5.9

Classification consistency

Each script included in the TAF grading was seen by between eight and 11 participants. Krippendorff's alpha ($K-\alpha$) was calculated using the nine fine grades and, separately, using the less granular KS2 grades. Double and triple moderated grades were also randomly generated for the TAF data to provide an understanding of the change in classification consistency with multiple views of a script. The same $K-\alpha$ values were calculated for the CJ grading.

Table 5 shows the classification consistency for CJ and TAF grading. $K-\alpha$ is higher for the fine grades than for KS2 grades because chance agreement is more likely in the latter.

Table 5 Classification consistency for CJ and TAF grading ($K-\alpha$ values)

		Fine grades		KS2 grades	
		$K-\alpha$	95% CI	$K-\alpha$	95% CI
TAF	Single moderation	0.60	(0.57-0.62)	0.48	(0.38-0.58)
	Double moderation	0.78	(0.76-0.80)	0.62	(0.48-0.71)
	Triple moderation	0.81	(0.79-0.83)	0.69	(0.57-0.78)
	Quadruple moderation	0.87	(0.85-0.88)	0.73	(0.59-0.85)
CJ		0.84	(0.82-0.86)	0.72	(0.61-0.80)

The figures show that classification consistency for CJ is higher than single, double and even triple moderation. So, although the resources required for CJ are may be equivalent to double moderation, the probability that a script will be given the same grade under repeated administrations is greater than triple moderation.

Figure 2 provides a heat map of classification consistency for the two approaches. It shows the grades awarded on two independent replications of the moderation or judging and is simulated from ten random samples of the study data. The darker the colour on the graph, the more often a combination of events occurs. The numbers within each square represent the probability of that event. The leading diagonal in each panel of the graph is where there is agreement between the grade awarded in each replication. The squares on (and close to) this leading diagonal are darker for CJ than for the TAF grading. This signifies a higher probability that CJ scripts will be given the same grade under repeated administrations. For example, under CJ, the probability that any script will receive an EXS grade if seen by different panels of judges is 0.42. The probability that a second set of judges would give this script a WTS grade is 0.07. Under the TAF, the probability that any script will receive two EXS grades if seen by two different moderators is 0.30. The probability that a second moderator would give this script a WTS grade is 0.15. The lower probabilities at GDS are simply a reflection of the timing of the study rather than a lack of classification consistency. As described earlier, very few pieces of work were deemed worthy of the highest grades.

Figure 2 A heatmap of the classification consistency for grades derived using CJ and the TAF

Discussion

This study set out to show the extent to which CJ and TAF approaches to the grading of writing measure the same construct, and the relative classification accuracy and consistency of each approach. While the TAF is highly specified in terms of what it asks teachers to reward in writing, it would appear that the approach of asking for ‘the better writing’ assessed using CJ rewards similar qualities (Figure 1). If it were possible to obtain reliable measures for the TAF grading it might then be feasible to discover the nuances of what is rewarded, but the poor classification accuracy and consistency of this approach undermines its specificity.

The poor reliability of the TAF grading is highlighted in its low classification consistency. With only 3 grades available a large degree of consistency would be expected by chance alone. It is clear that beyond this chance level of agreement, the TAF is not leading to reliable, repeatable decisions about the quality of writing. Breaking the grading down into nine fine grades, which minimises the level of agreement by chance, the CJ approach has twice the classification accuracy of a TAF approach.

The results reported here are not surprising. Firstly, the poor reliability of the TAF approach is supported by marking reliability studies (for example, He et al., 2013; Rhead et al., 2018). Secondly the improvement of CJ approaches over rubric approaches is supported by studies which have compared the extent to which each predict results in similar constructs (Benton & Gallacher, 2018; Holmes et al., 2017). Finally, the advantage of CJ is theoretically predictable! A CJ approach has the advantage not only of a simple decision mechanism but also of multiple judges combining views via a powerful statistical model (Benton & Gallacher, 2018).

Where this research does depart from other studies into CJ is that the efficiency is found to be high. On average grading a single script took 4 minutes, which seems a worthwhile investment in a high-stakes context. Finally, no bias was found by gender or socio-economic status in switching from a TAF approach to a CJ approach.

The major limitation of the study is that the pieces considered were short and produced under controlled conditions. Many would argue that the process of revision is critical to developing good writing skills, and this process was not assessed here. There is absolutely no reason to believe, however, that the same results would not play out were the writing to have been produced in less controlled conditions with opportunities for revision. If the ability to successfully revise work is valued, then an assessment could easily be created that allows for revision, for writing in different genres and for different audiences. Quibbles over what exactly is assessed, however, should always be considered in the light of the reliability with which they can be assessed. This study suggests that the poor reliability of any rubric-based approach to assessing writing undermines its specificity in objectives.

Acknowledgements

We would like to thank our participating schools & pupils, our senior moderator and our teacher markers. We are also grateful to NESTA and the DFE for funding this project from their EdTech Innovation Fund which supports more effective use of technology in education.

References

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561–573.
- Andrich, D. (1982). An index of person separation in Latent Trait Theory, the traditional KR-20 Index, and the Guttman Scale Response Pattern. *Education Research and Perspectives*, 9(1), 95–104.
- Benton, T., & Gallacher, T. (2018). Is comparative judgement just a quick form of multiple marking? *Research Matters: A Cambridge Assessment Publication*, 26, 22–28.
- Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4), 324.
- Bramley, T., & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy and Practice*, 26(1).
- Cuff, B. M. P., Howard, E., Mead, R., & Newton, P. E. (2018). *Key stage 2 writing moderation: Observations on the consistency of moderator judgements*.
<https://www.gov.uk/government/publications/observations-on-the-consistency-of-moderator-judgements>
- Hanson, B. A. ., & Brennan, R. L. . (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, 27(4), 345–359.
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1), 77–89.
- He, Q., Anwyll, S., Glanville, M., & Deavall, A. (2013). An investigation of the reliability of marking of the Key Stage 2 National Curriculum English writing tests in England. *Educational Research*, 55(4), 393–410.
- 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.
- Holmes, S., Black, B., & Morin, C. (2017). *Marking reliability studies 2017: Rank ordering versus marking – which is more reliable?*
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/859250/Marking_reliability_-_FINAL64494.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. *Australian Educational Researcher*, 42(4).
- Lee, W. C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, 26(4), 412–432.
- Luce, R. D. (1959). *Individual choice behaviours: A theoretical analysis*. NY: J Wiley.
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18–39.
- Rhead, S., Black, B., & Pinot de Moira, A. (2018). *Marking consistency metrics: An update*.

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/759207/Marking_consistency_metrics_-_an_update_-_FINAL64492.pdf

Standards & Testing Agency. (2019). *Key stage 2 teacher assessment guidance: For schools and local authorities*.

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/849018/2020_key_stage_2_teacher_assessment_guidance.pdf

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy and Practice*, 26(1), 59–74.

Wheadon, C. (2014). Classification accuracy and consistency under item response theory models using the package classify. *Journal of Statistical Software*, 56(10).

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy and Practice*, 27(1), 46–64.

Wheadon, C., & Pinot de Moira, A. (2013). Gains in marking reliability from item-level marking: is the sum of the parts better than the whole? *Educational Research and Evaluation*, 19(8), 665–679.