Comparative approaches to the assessment of writing: Reliability and validity of benchmark rating and comparative judgement

Renske Bouwer[1], Marije Lesterhuis[2], Fien De Smedt[3], Hilde Van Keer[3] & Sven De Maeyer[2]

[1] Utrecht University | the Netherlands

[2] University of Antwerp | Belgium

[3] Ghent University | Belgium

Abstract: In the past years, comparative assessment approaches have gained ground as a viable method to assess text quality. Instead of providing absolute scores to a text as in holistic or analytic scoring methods, raters in comparative assessments rate text quality by comparing texts either to pre-selected benchmarks representing different levels of writing quality (i.e., benchmark rating method) or by a series of pairwise comparisons to other texts in the sample (i.e., comparative judgement; CJ). In the present study, text quality scores from the benchmarking method and CJ are compared in terms of their reliability, convergent validity and scoring distribution. Results show that benchmark ratings and CJ-ratings were highly consistent and converged to the same construct of text quality. However, the distribution of benchmark ratings showed a central tendency. It is discussed how both methods can be integrated and used such that writing can be assessed reliably, validly, but also efficiently in both writing research and practice.

Keywords: writing assessment, benchmark rating, comparative judgement, reliability, convergent validity

journal of
WRITING RESEARCH

Bouwer, R., Lesterhuis, M., De Smedt, F., Van Keer, H., & De Maeyer, S. (2024). Comparative approaches to the assessment of writing: Reliability and validity of benchmark rating and comparative judgement, *Journal of Writing Research, 15*(3), 498-518. DOI: 10.17239/jowr-2024.15.03.03

Contact: Renske Bouwer, Utrecht University, Institute for Language Sciences, Trans 10, 3512 JK Utrecht | The Netherlands – r.bouwer@uu.nl – 0000-0003-0434-0224

1. Introduction

At the heart of writing education and research lies the assessment of writing. Without any adequate and appropriate measure of the quality of written texts, students fail to have insight in their ability to write, teachers cannot monitor and promote students' writing development in class, policymakers do not know whether students' writing skills meet the national standards, and researchers cannot make claims on the effectiveness of writing processes and interventions (cf. Van Steendam et al., 2012, pp. x-xxi). For all these stakeholders, however, it is a true challenge to assess writing quality in a reliable and valid manner, especially in low-stakes classroom contexts in which there are relatively many texts and few assessors. This study aims to extend our knowledge on writing assessment in small-scale studies by investigating the reliability and validity of scores based on two comparative assessment approaches that are increasingly being used in educational research and practice: benchmark ratings and comparative judgement. We will empirically investigate the extent to which they measure text quality in a consistent and appropriate manner. This knowledge is of vital importance for educators and researchers to justify a valid and reliable interpretation of text quality ratings based on a comparative assessment of writing, and whether both comparative approaches can be used interchangeably.

That assessing writing performance is notoriously difficult is mainly due to the complexity and multidimensionality of text quality (Huot, 1990). A text cannot be considered as either good *or* bad; on the contrary it can be simultaneously good *and* bad, for different reasons. For instance, a text can be well-written because of its linguistic features, i.e., with grammatically correct sentences, diverse and complex words and connectives, and free of spelling and punctuation errors. The same text, however, can also be considered as poorly written when it fails to fulfil its communicative purpose. For instance, when ideas or arguments in the text are not yet well-developed or logically structured. Consequently, when aiming at making evaluative judgements of students' quality of writing beyond the linguistic level, which is often the aim in educational research and practice, assessors need to take the communicative and content aspects of writing into account. But doing so in a reliable and valid manner is a challenge.

When assessors are completely free in forming a holistic impression of text quality, their scores may vary to a large extent. One of the reasons for the large rater variance is that assessors have their own standards of what constitutes a good text and focus on different criteria when evaluating text quality (Myford & Wolfe, 2003). Another reason is that they vary in how they use a scoring scale, with some assessors who are more severe or lenient than others (Leckie & Baird, 2011). These differences between assessors affect the reliability and, hence, the validity of text quality scores (Bouwer et al., 2023). To ensure that assessors take all relevant aspects into account when evaluating text quality and use the scoring scale consistently, assessors need clear rating instructions including examples and predefined rating criteria, such as content, structure, style, grammar, mechanics (cf. Weigle, 2002).

The freedom of assessors can be even more restricted by analytic rating schemes, such as criteria-lists, rubrics, or checklists (Jonsson & Svingby, 2007). In these analytic methods, assessors score predefined criteria separately. These analytic scores can then be combined into one composite score for text quality. A meta-analysis by Jonsson and Svingby (2007) shows that analytic and holistic rubrics combined with rater training are associated with acceptable levels of score reliability. However, there are also studies demonstrating that even under these more restricted conditions, assessors remain to vary in how they interpret and apply analytic criteria to a text (Barkaoui, 2011; Lumley, 2002). Moreover, it can be questioned whether the sum of separately scored sub-dimensions are a valid representation of overall text quality and whether text features can be considered as independent constructs at all (Huot, 1990; Mabry, 1999; Sadler, 2009). In sum, evaluating text quality seems to be a challenge and neither a holistic nor an analytic assessment method guarantees a valid and reliable interpretation and use of writing scores.

2. Comparative Approaches to the Assessment of Writing

There Recently, a comparative approach to the assessment of writing has been introduced as a promising alternative for researchers and educators, also in low stakes writing performance settings (Bouwer et al., 2023; Heldsinger & Humphry, 2010; Lesterhuis et al., 2017; Pollitt, 2004, 2012). Instead of evaluating one text after the other as is the case in holistic and analytic methods, a comparative approach allows assessors to evaluate text quality by comparing texts to each other and to determine which one is of higher quality. This is considered to be a more natural way of how we make evaluative judgements. Laming (2004) states that every judgement is actually a comparison, either with an internal standard, or with work that has been seen before. This may explain the well-known order effects in writing assessment (Myford & Wolfe, 2003; Weigle, 2002), in which texts of average quality receive a relatively higher score for text quality if they are evaluated after a series of low-quality texts than after a series of high-quality texts. By comparing texts directly to each other, it is assumed that texts are no longer implicitly compared to previously assessed texts or to an unknown internal standard.

Two procedures adopting a comparative approach to the assessment of writing are currently prevailing: benchmark ratings and comparative judgements (CJ). Following a benchmark rating procedure, raters compare each text to prototypical texts (i.e., benchmarks) on a continuous rating scale and score the text accordingly. The benchmarks represent different levels of text quality, ranging from very poor to very good. For each benchmark it is described why it is of higher (or lower) quality than the other benchmarks. This description can include holistic information on the overall quality of the text as well as more detailed and analytic information for specific criteria. For examples of studies that have used such a benchmark rating procedure, see Bouwer et al. (2018) or Schoonen (2005).

In CJ, texts are not compared to benchmarks but directly to each other in a series of pairwise comparisons. This implies that assessors do not provide scores, but they only have to indicate in each pairwise comparison which of the two texts they consider the one with the highest quality. This procedure eliminates score variance due to differences between raters in how they use the rating scale (e.g., effects of central tendency or rater severity/lenience, see Myford & Wolfe, 2003). Based upon the probability that a text is selected as the best in each pair, texts can be ranked on a scale from low to high quality (Thurstone, 1927). Applying the Bradley-Terry-Luce model to all these pairwise comparisons also result in scores for text quality (in logit values), which can be considered as the shared consensus of assessors regarding the extent to which a particular text is of higher quality than the other texts (Van Daal et al., 2019). For examples of studies that have implemented the method of CJ, see Van Daal et al. (2022).

Research has consistently shown that both comparative assessment procedures increase the reliability and validity of text quality ratings. Benchmark ratings, for instance, are associated with less rater variance than holistic ratings, and with less task-specific rater variance than analytic ratings (Bouwer & Koster, 2016; Schoonen, 2005; Van den Bergh et al., 2012). This improves the generalizability of benchmark ratings to students' overall writing performance, which implies that the benchmarking method leads to not only reliable but also valid ratings. Blok (1986) reported that the method of benchmark ratings is also a feasible method, as instructions for providing scores of text quality by comparing texts to a benchmark were easily applied by raters. The method has been implemented in several writing research studies for the assessment of text quality, with inter-rater reliability levels ranging from .72 to .90 (Bouwer & Koster, 2016; De Milliano et al., 2012; De Smedt & Van Keer, 2018; Schoonen, 2012; Tillema et al., 2012).

Comparative judgement is also associated with highly reliable ratings for text quality, given that sufficient comparisons per text are made (Bramley, 2015; Pollitt, 2012; Verhavert et al., 2017). A recent meta-analysis of 49 CJ assessments revealed that on average, when pairs are randomly selected, 12 comparisons were needed to reach a minimum level of reliability of .70, and 17 comparisons for a reliability of at least .80 (Verhavert et al., 2019). Despite the workload, the average judgement time per text in a comparison is still more than 5 times lower than when the same text is judged in an absolute way (Coertjens et al., 2017). Furthermore, the total number of comparisons can be distributed among multiple raters, which decreases the workload per individual rater. Researchers have also shown that including multiple raters in a CJ session will increase the validity of the results, as the shared consensus of multiple raters appears to be a good reflection of all relevant aspects of the writing task (Jones & Alcock, 2014; Lesterhuis, 2018; Van Daal et al., 2019).

3. Present Study

Based upon previous research findings we can conclude that comparative assessment approaches can lead to reliable and valid ratings of text quality. However, while the two above-mentioned comparative approaches are similar, they do differ on some key aspects. For instance, they differ in the extent to which assessors are supported during the evaluative process. In the benchmark procedure, assessors are provided with standards and descriptions for high and low text quality that should assist them with scoring consistently over time. These standards are, however, not present in CJ. This might decrease the reliability of the scores in CJ in comparison to benchmark ratings, even though assessors in CJ can make a relatively easier decision (i.e., only pointing out the best text in a pair, instead of also assigning scores to the texts). In addition, even though results regarding the effects of rater training for the reliability of performance assessments are generally mixed (Lumley & McNamara, 1995; Rezaei & Lovorn, 2010), and assessors appear to focus predominantly on aspects that are related to the quality of the argumentation and organisation when comparing texts (Lesterhuis, 2018), it is yet unknown whether a training in CJ is required for assessors to keep their judgements consistent, and hence reliable, over time.

Comparable levels of reliability are not sufficient to conclude that both methods are interchangeable, it is also needed to get insight into the convergent validity, that is, whether the scores in both procedures converge. High correlations between scores indicate that both procedures measure the same underlying construct (Messick, 1989). Furthermore, it can be questioned whether the scoring distribution is comparable in both methods. For optimal use in educational practice, text quality scores have to follow a normal distribution with sufficient score variation and no central tendency (Borsboom et al., 2004; Mabry, 1999). However, assessment research has shown that assessors can use rating scales quite differently, such as only using the central range of the scale instead of also the extreme scores (i.e., central tendency or restriction-of-range, Humphry & Heldsinger, 2014; Myford & Wolfe, 2003). While assessors in the CJ procedure only have to indicate what they consider to be the best text in a pair, assessors in the benchmarking procedure have to provide scores by using the benchmark rating scale, which leaves room for individual rater biases. This may negatively affect the scoring distribution in the benchmark rating procedure, making it difficult to adequately discriminate between texts of high and low quality.

The aim of the present study is to evaluate and compare the reliability and validity of benchmarking and CJ. This knowledge will support researchers and educators to make well-informed decisions on how to best implement which comparative assessment procedure in what context. More particularly, this study focuses on the following research questions:

1. To what extent do benchmarking and CJ ratings, either by trained or untrained raters, lead to comparable levels of reliability?
2. To what extent do ratings obtained by the benchmark and CJ rating procedure converge to the same construct of writing quality as reflected by their correlation (i.e., convergent validity)?
3. To what extent do benchmark and CJ ratings have a comparable scoring distribution?

In line with previous research on comparative assessment (Bouwer & Koster, 2016; Lesterhuis et al., 2017; Pollitt, 2012; Verhavert et al., 2019), we expected high reliabilities for both comparative assessment methods, regardless of whether assessors in the CJ method were trained or not. Moreover, as both methods are based on a comparative approach, we expected the ratings of CJ to converge with the benchmark ratings. Finally, we expected that assessors are able to differentiate between texts of low and high quality in both rating procedures, but that CJ scores will be better distributed over the entire scale, given the potential influence of individual scoring biases in the benchmark ratings (cf. Myford & Wolfe, 2003).

4. Method

4.1 Participants

A total of fourteen undergraduate students (thirteen female, one male), studying at the Department of Educational Studies, participated as raters in this study. Their mean age was 22 years (*SD* = 1.15). None of the participating students had any experience with one of the rating procedures, nor with rating the quality of texts written by primary school students. Students were randomly assigned to one of the three rating conditions: Two students were assigned to the benchmark rating procedure, six students were assigned to the CJ rating procedure with training, and six students were assigned to the CJ rating procedure without training. We assigned more raters to the CJ conditions to ensure that each rater received a comparable number of comparisons in each condition (183 individual comparisons for benchmark ratings versus 225 comparisons for CJ), and hence, invested an equal amount of rating time. A recent meta-analysis revealed that the reliability level of CJ is affected by the number of comparisons per rater, but not by the overall number of raters in the assessment (Verhavert et al., 2019).

4.2 Materials and procedure

4.2.1 Text sample

Raters assessed the quality of 183 descriptive texts written by fifth ($n = 97$) and sixth-graders ($n = 86$) from nine primary schools. The texts were randomly selected from a larger research project on the state-of-the-art of descriptive and narrative writing in primary education in Belgium (De Smedt et al., 2015, 2017). The writing task included a visual prompt of an alien and a school building with the following instruction: "Describe to the alien what a school is", see Appendix A. Students received the writing task in class by trained research assistants, and they had 40 minutes to individually write the text. To control for presentation effects in the ratings of text quality, the handwritten texts were typed, and spelling, punctuation, and capital errors were corrected (cf. Graham et al., 2011; De Smedt et al., 2017).

4.2.2 Benchmark ratings

Participants who were assigned to the benchmark rating procedure were asked to rate each text holistically by comparing it to five benchmark texts that represented different levels of text quality on a continuous interval scale. The benchmarks originated from the same research project as the text samples in the present study (De Smedt et al., 2015, 2017). Appendix B provides an English translation of the benchmark scale including benchmark descriptions. Benchmarks were selected based on pre-ratings of two independent expert raters. There were two criteria for benchmarks to be selected: (1) they had to be a good representative for one of the five scale points, and (2) there had to be high agreement between the expert raters for the quality of that particular text (for a more detailed explanation of the selection procedure, see De Smedt et al., 2017). The average benchmark text received an arbitrary score of 100 points and there was an interval of 15 points between the benchmarks that represented lower or higher levels of text quality. The text with the lowest and highest quality were indicated by a score of 70 and 130 points respectively. Raters were, however, allowed to use the whole rating scale ranging from 0 to infinite, including all possible scores in-between benchmarks. For each benchmark, strong and weak points were explained according to specific criteria, i.e., genre conventions, idea development, text structure, sentence structure, and word choice. Raters were explicitly instructed to take these criteria into consideration when assessing the texts using the benchmark scale, but to refrain from making analytic evaluations in an absolute way. All grammar and spelling errors were eliminated from the selected benchmarks, to ensure that raters' attention was not drawn away to mechanics.

Both raters assigned to the benchmark rating procedure in the present study received a short training session of half an hour in advance on how to use the benchmark scale. This training consisted of an explanation of the comparative assessment in general and the benchmark rating procedure specifically. They also discussed the writing prompt, the accompanying criteria for text quality and the benchmarks that represented performance levels on the benchmark rating scale. In addition, raters practised and discussed the benchmark rating procedure through rating a couple of example texts in a holistic manner by comparing them to the benchmarks on the rating scale. All the training materials were provided on paper. After the training, the raters independently rated all student texts.

4.2.3 Pairwise comparisons

In the comparative judgement condition texts were randomly presented in pairs to raters on a computer screen. To optimise the level of reliability of the comparative judgements, each text was compared fifteen times to random texts within the sample (Verhavert et al., 2019). This resulted in a total of 1353 comparisons for both the training and non-training condition, which were randomly distributed over the six raters in each condition, such that each rater made 225 comparisons. To automatize pair selection and the distribution over raters, we used the Digital Platform for the Assessment of Competences (D-PAC).

Raters in the non-training condition received short instructions on paper on how to log into D-PAC at home and how to make comparisons. They were also provided with the writing prompt and the same assessment criteria as in the benchmark condition (see Appendix B): genre conventions, idea development, text structure, sentence structure, and word choice. These criteria were also accessible online during the assessment through a button on the screen. They were explicitly instructed to take the criteria into consideration when making holistic and pairwise comparisons, but to refrain from making analytic evaluations of individual texts. They did not receive any specific training concerning these criteria and completed the pairwise comparisons individually at home, at their own pace.

Raters in the training condition received a short group-training session of half an hour that was comparable to the training for raters in the benchmark condition. The purpose of this training was to inform them on the idea behind comparative assessment and to discuss the writing prompt and criteria for text quality. Second, raters practised with making comparative judgements by comparing a couple of example texts on paper. We used the benchmark texts for this purpose, as this allowed us to resemble the training information provided in the benchmark condition. After the training, the raters independently finished all assigned pairwise comparisons online at their own pace. They were able to take as many breaks as they needed during the assessment. On average, they needed four hours to complete all 225 comparisons, which was comparable to the non-training and benchmarking condition.

4.3 Data analysis

We estimated the reliability of the benchmark ratings and the comparative judgements (RQ1), in order to compare how consistently raters evaluate text quality in the three rating conditions. For the benchmark procedure, the inter-rater reliability was estimated by the Intraclass Correlation Coefficient (ICC) using the two-way random effects model for consistency of both single and multiple raters (k = 2). The ICC for multiple raters reflects the reliability of average scores across both raters. As these ratings are measured at interval level at best (Suppes & Zinnes, 1963), we used raters' standardised scores to control for absolute differences in scores due to individual harshness or leniency.

In the CJ procedure, the reliability was estimated by Separation Scale Reliability (SSR). The SSR is calculated by applying the Bradley-Terry-Luce model on all pairwise comparisons that were made. This model estimates logit scores for each text based on the number of times a particular text was chosen as the better text of a pair, while taking into account the quality of the texts they were compared to. In this way, the logit scores represent differences in text quality, which can be used for further data analyses. The SSR is an indication of the consistency of the logit scores, and therefore represents the inter-rater reliability (Verhavert et al., 2017). The SSR is comparable to the ICC for multiple raters, as they both reflect the reliability of average scores across raters (Bramley, 2015; Gwet, 2014; Verhavert et al., 2017).

The convergent validity (RQ2) is estimated by the Pearson correlation between the CJ and benchmark scores. A significant, positive correlation coefficient of at least .70 indicates that the scores converge and, hence, that both rating procedures measure the same construct (Cook & Campbell, 1979; Post, 2016).

To compare the score distribution between both rating procedures (RQ3), we executed a standardisation procedure of the CJ scores, taking the same intervals as used within the benchmark procedure (M=100; SD=15). This procedure also allowed us to examine whether the benchmark texts were equally distributed among the CJ scale. The Shapiro-Wilk test of normality was used to investigate whether the standardised scores were normally distributed. We also calculated the skewness and kurtosis of the scores, for which estimates deviating from zero respectively indicate the degree of asymmetry and peakedness in the scoring distribution.

5. Results

5.1 Reliability

The reliability of the benchmark ratings was high, with an ICC of .91 for average scores across both raters and an ICC of .84 for single raters. This indicates that raters are highly consistent in the text quality scores they provide when using a benchmark rating scale.

The reliability of the text quality scores provided in the CJ condition were high as well, but somewhat lower than the scores in the benchmark rating condition, with an SSR of .82 for the trained raters and .79 for the non-trained raters. As the reliability of CJ is strongly affected by the number of pairwise comparisons (Verhavert et al., 2019), we estimated the evolution of the SSR for the trained and non-trained raters as a function of the number of comparisons per text (see Figure 1). For the trained raters nine comparisons per text were needed to reach a reliability level of .70, in comparison to the non-trained raters who needed to make ten comparisons per text for an equal reliability level. It also shows that the small difference in the consistency of trained and non-trained raters becomes negligible when a reliability level of .80 is aimed at.

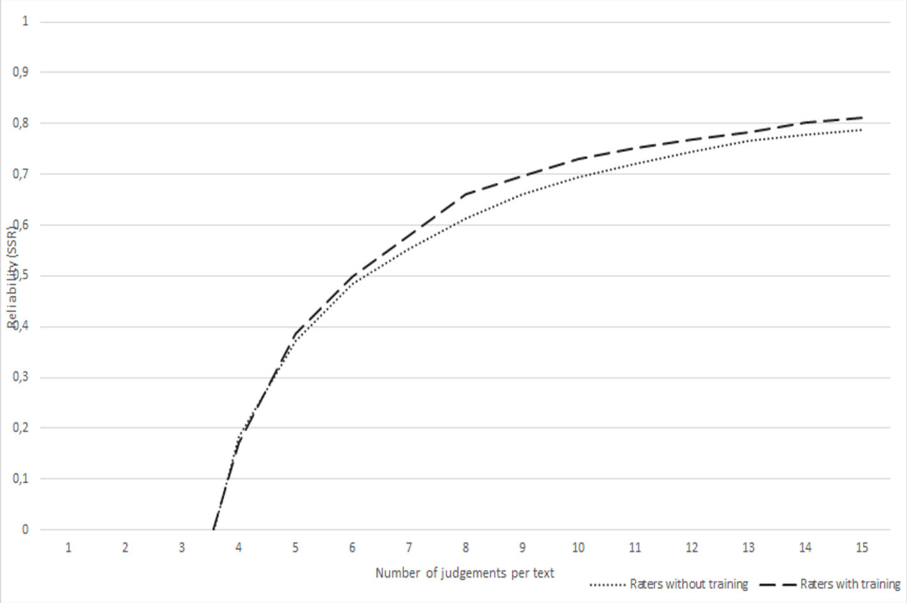

Figure 1: *Level of reliability (SSR) in the CJ condition as a function of the number of judgements per text*

5.2 Convergent Validity

Results show a positive and strong correlation between the text quality scores in the benchmark condition and the CJ condition for trained raters, $r = .87$, $p < .001$. The same was demonstrated for scores between the benchmark condition and the CJ condition with non-trained raters, albeit somewhat lower, $r = .81$, $p < .001$. This indicates that scores obtained either by comparing texts to benchmarks or by pairwise comparisons largely converge to the same construct of writing quality.

5.3 Score Distribution

In order to compare the distribution of CJ scores to the benchmark scores, we standardized and converted the CJ logit scores to a comparable interval rating scale with a mean of 100 and standard deviation of 15. The Shapiro-Wilk test showed that while the CJ scores were found to be normally distributed (*W*(178) = .99, *p* = .39), the benchmark scores were not (*W*(178) = .92, *p* < .001). Figure 2 shows the boxplots for both rating procedures, demonstrating that the benchmark ratings are negatively skewed (-1.13, *SE* = .18), due to a handful of texts that were rated lower than would be expected based on a normal distribution. As a result, the scoring range of 80 (min = 50 and max = 130) for the benchmark ratings is larger than the scoring range for the CJ ratings, which is 71 (min = 64 and max = 135). However, the interquartile range of the benchmark ratings is smaller than that of the CJ ratings. This suggests that there is a stronger central tendency for the benchmark ratings than for the CJ ratings. This is also illustrated by a positive kurtosis of 1.43 (*SE* = .36) for benchmark scores, in comparison to a non-significant kurtosis of -0.22 (*SE* = .36) for CJ scores.

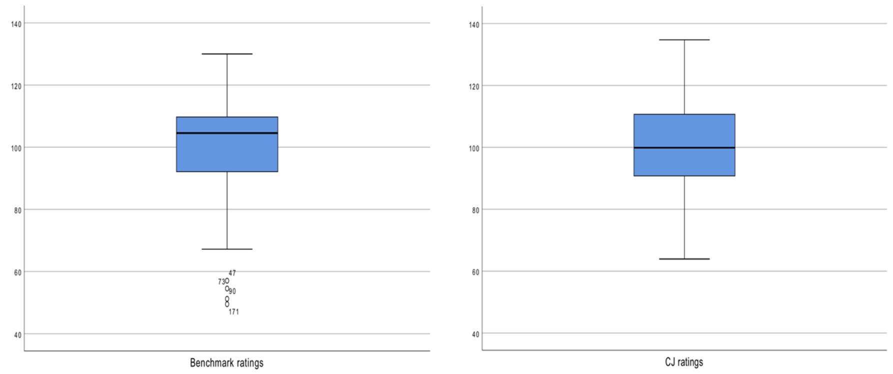

Figure 2: Boxplot of score distribution in the benchmark rating condition (left) and the CJ training condition (right)

As the texts in the CJ rating procedure also included the benchmark texts, we were able to compare the scoring distribution in more detail. Table 1 shows that the benchmarks were rank-ordered in a similar way in the CJ condition, with the benchmark of the lowest and highest quality being scored as respectively the lowest and highest in the CJ condition as well. Further, the benchmark of average quality, that was provided with an arbitrary score of 100 points in the benchmark rating procedure, received a similar score of 98 points in the CJ condition. In both rating procedures, an average of forty percent of the texts received a lower score than the average benchmark. However, for the benchmarks of low and high quality the absolute and percentile scores were different between conditions. In the benchmark condition most scores were located around the midpoint of the rating scale, whereas the scores for the CJ condition were more dispersed.

Table 1. Absolute and Percentile Scores of Benchmarks in the Benchmark and CJ Condition

	Benchmark condition		CJ condition	
Benchmark text	Score	Percentile score	Score	Percentile score
Highest quality	130	100	121	93
High quality	115	90	111	76
Average quality	100	40	98	43
Low quality	85	14	94	33
Lowest quality	70	5	74	5

6. Discussion

In this study we evaluated and compared text quality scores based on two comparative rating procedures for the assessment of writing: benchmarking and CJ. Results demonstrated that both rating procedures were associated with high levels of reliability, and that the benchmark scores converged highly with the scores obtained in the CJ procedure. The reliability and convergent validity were somewhat higher for the scores in the training CJ condition than in the non-training CJ condition. We also found that the benchmarks in the CJ procedure were ranked in the same order as in the benchmark rating scale. In sum, the results indicate that both the benchmark and CJ rating procedure measure the same underlying construct with the same consistency across raters, and hence, are both valid and reliable methods to rate text quality.

Regarding the distribution of ratings, however, it was shown that the ratings in the CJ procedure were more dispersed and normally distributed than the benchmark ratings. In particular, benchmark ratings were more negatively skewed and raters in the benchmark condition generally showed a stronger central tendency, i.e., scoring closer to the midpoint of the scale, resulting in a positive kurtosis. This suggests that comparative judgement promotes the differentiation between texts of low and very low quality, as well as between high and very high quality, which increases the validity of the CJ scores (cf. Borsboom et al., 2004). This finding is especially important when text quality scores are used in an absolute instead of a relative way, for instance, by making summative decisions based on absolute standards (e.g., passing a course if the score is 6 or higher).

To validate the findings of this study, replication with a different sample of texts in another context is needed. For instance, longer texts written by students in higher education may increase the complexity of the rating task, which could have consequences for the reliability of the ratings. However, a recent meta-analysis on comparative judgement demonstrates that the format of the task (e.g., texts, images, video, or portfolios) hardly affects the level of reliability (Verhavert et al., 2019). It is also recommended to replicate the study with another sample of more experienced raters, even though previous studies showed no differences between student raters and more experienced raters for comparative judgement (Lesterhuis, 2018).

In practice, reliable and valid interpretations of writing scores are not the only criteria for deciding to use a rating procedure, it also needs to be efficient and feasible to apply (cf. Messick, 1989). This is particularly relevant for low-stakes assessments in educational or research contexts when only a few assessors are available. If efficiency and feasibility are taken into account, the benchmark rating procedure seems to be more cost-efficient than CJ. In the benchmark rating procedure not only less raters were needed for comparable levels of reliability (2 instead of 6), but they, as a group, also needed less evaluations than the raters in the CJ condition to obtain reliable scores. This observation is in line with previous research in other contexts, showing that the total assessment time that is needed to obtain reliable scores in CJ easily exceeds the time needed for reliable absolute ratings (Coertjens et al., 2017; Goossens & De Maeyer, 2018). Although the process of pairwise comparisons is quite easy and fast for raters to apply (Laming, 2004; Pollitt, 2004), even without training, the increase in time investment that is needed for CJ decreases its feasibility in educational practice.

Even though benchmark ratings seem to be more efficient, especially when only a limited number of raters is available, valid interpretation and use of benchmark ratings depends heavily on the quality of the benchmark scales. In this study we used an already available benchmark scale. However, this benchmark cannot be used for all writing tasks: it is genre-specific (i.e., descriptive texts) and task-specific (i.e., explaining what a school is). The development of a new benchmark scale requires ample time and expertise. For instance, a representative set of texts is needed in order to select benchmarks that provide adequate support for rating text quality across the whole rating scale. Benchmarks that do not adequately distinguish between low, average, and high performance of writing, have a negative impact on the quality of the benchmark ratings (Osborn Popp, Ryan, & Thompson, 2009).

To select benchmarks in an efficient and adequate way, researchers have proposed to integrate both comparative rating procedures in a two-stage process (McGrane, Humphry, & Heldsinger, 2018; Heldsinger & Humphry, 2013; Lesterhuis et al., 2017). In stage 1, the results of a CJ rating session are used for the calibration and selection of benchmarks of different performance levels resulting in a benchmark rating scale. In stage 2, this benchmark rating scale is used by individual raters to assess a second set of texts. This allows teachers to rate writing quality in a valid and reliable way (Heldsinger & Humphry, 2013; for an example, see De Smedt et al., 2020).

In conclusion, this study has demonstrated that both CJ and a benchmark rating procedure lead to reliable and comparable scores of text quality, which can be validly interpreted and used as indicators for text quality by both researchers and educators. The benefit of using a comparative approach of rating text quality is that raters can assess the holistic quality of a text, rather than breaking up the assessment in parts and rating only subdimensions of writing quality. By doing so, a comparative approach to the assessment of writing quality is doing justice to the overall quality of the text.

References

Barkaoui, K. (2011). Effects of Marking Method and Rater Experience on ESL Essay Scores and Rater Performance. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 279–293. http://doi.org/10.1080/0969594X.2010.526585

Blok, H. (1986). Essay rating by the comparison method. *Tijdschrift Voor Onderwijsresearch*, *11*, 169–176.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071. https://doi.org/10.1037/0033-295x.111.4.1061

Bouwer, R., & Koster, M. (2016). *Bringing writing research into the classroom. The effectiveness of Tekster, a newly developed writing program for elementary students* (Unpublished doctoral dissertation). Utrecht University.

Bouwer, R., Koster, M., & Van den Bergh, H. (2018). Effects of a strategy-focused instructional program on the writing quality of upper elementary students in the Netherlands. *Journal of Educational Psychology, 110*(1), 58-71. http://doi.org/10.1037/edu0000206

Bouwer, R., Van Steendam, E., & Lesterhuis, M. (2023). Assessing writing performance: Guidelines for the validation of writing assessment in intervention studies. In F. De Smedt, R. Bouwer, T. Limpo, & S. Graham (Eds.), *Conceptualizing, designing, implementing, and evaluating writing interventions*. Brill.

Bramley, T. (2015). *Investigating the reliability of Adaptive Comparative Judgment* (pp. 1-17). Cambridge Assessment.

Coertjens, L., Lesterhuis, M., Verhavert, S., Van Gasse, R., & De Maeyer, S. (2017). Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: een afweging van betrouwbaarheid en tijdsinvestering. *Pedagogische Studiën*, *94*(4), 283–303.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: design & analysis issues for field settings.* Houghton Mifflin.

Van Daal, T., Lesterhuis, M., De Maeyer, S., & Bouwer, R. (2022). Validity, reliability, and efficiency of comparative judgement to assess student work. *Frontiers in Education*, *7*:1100095.
http://doi.org/10.3389/feduc.2022.1100095

De Milliano, I., van Gelderen, A., & Sleegers, P. (2012). Patterns of cognitive self-regulation of adolescent struggling writers. *Written Communication, 29*(3), 302-325. http://doi.org/10.1177/0741088312450275

De Smedt, F., Graham, S., & Van Keer, H. (2020). 'It takes two': the added value of structured peer-assisted writing in explicit writing instruction. *Contemporary Educational Psychology, 60*, 101835. https://doi.org/10.1016/j.cedpsych.2019.101835

De Smedt, F., Merchie, E., Barendse, M., Rosseel, Y., De Naeghel, J., & Van Keer, H. (2017). Cognitive and motivational challenges in writing: studying the relation with writing performance across students' gender and achievement level. *Reading Research Quarterly, 53*(2), 249–272.
http://doi.org/10.1002/rrq.193

De Smedt, F., & Van Keer, H. (2018). Fostering writing in upper primary grades: a study into the distinct and combines impact of explicit instruction and peer assistance. *Reading and Writing, 31*(2), 325-354. http://doi.org/10.1007/s11145-015-9590-z

De Smedt, F., Van Keer, H., & Merchie, E. (2015). Student, teacher and class-level correlates of Flemish late elementary school children's writing performance. *Reading and Writing, 29*(5), 1–36. http://doi.org/10.1007/s11145-015-9590-z

Goossens, M., De Maeyer, S. (2018). How to Obtain Efficient High Reliabilities in Assessing Texts: Rubrics vs Comparative Judgement. In E. Ras, & A. Guerrero Roldán (Eds.), *Technology Enhanced Assessment. Communications in Computer and Information Science*, vol. 829. Springer.
https://doi.org/10.1007/978-3-319-97807-9_2

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, *37*(2), 1–20. https://doi.org/10.1007/bf03216919

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.
http://doi.org/10.1080/00131881.2013.825159

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. https://doi.org/10.3102/0013189x14542154

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*(2), 201–213. http://doi.org/10.2307/358160

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, *39*(10), 1774–1787. https://doi.org/10.1080/03075079.2013.821974

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144.
https://doi.org/10.1016/j.edurev.2007.05.002

Laming, D. (2004). Marking university examinations: some lessons from psychophysics. *Psychology Learning & Teaching*, *3*(2), 89–96. http://doi.org/10.2304/plat.2003.3.2.89

Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418.

Lesterhuis, M. (2018). *The validity of comparative judgement for assessing text quality. An assessor's perspective* (Unpublished doctoral dissertation). University of Antwerp.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Comparative Judgement as a promising alternative. In E. Cano & G. Ion (Eds.), *Innovative practices for higher education assessment and measurement* (pp. 119–138). IGI Global. http://doi.org/10.4018/978-1-5225-0531-0

Lumley, T. (2002) Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3)*,* 246-274. https://doi.org/10.1191/0265532202lt230oa

Lumley, T., & McNamara, T. F. (1995). Rater Characteristics and Rater Bias: Implications for Training. *Language Testing*, *12*(1), 54–71. https://doi.org/10.1177/026553229501200104

Mabry, L. (1999). Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment. *The Phi Delta Kappan, 80*(9), 673-679.

McGrane, J. A., Humphry, S. M., & Heldsinger, S. (2018). Applying a thurstonian, two-stage method in the standardized assessment of writing. *Applied Measurement in Education, 31*(4), 297-311.
http://doi.org/10.1080/08957347.2018.1495216

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). American Council on Education and National Council on Measurement in Education.

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.

Osborn Popp, S. E., Ryan, J. M., & Thompson, M. S. (2009). The critical role of anchor paper selection in writing assessment. *Applied Measurement in Education*, *22*(3), 255–271. http://doi.org/10.1080/08957340902984026

Pollitt, A. (200, June). *Let's stop marking exams* [Paper presentation]. IAEA Conference. Philadelphia.

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 281–300. http://doi.org/10.1080/0969594X.2012.665354

Post, M. W. (2016). What to do with "moderate" reliability and validity coefficients. *Archives of Physical Medicine and Rehabilitation, 97*, 1051-10522. https://doi.org/10.1016/j.apmr.2016.04.001

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*(1), 18–39. http://doi.org/10.1016/j.asw.2010.01.003

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159-179. http://doi.org/10.1080/02602930801956059

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1–30. http://doi.org/10.1191/0265532205lt295oa

Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Bergh (Eds.), *Measuring Writing: Recent Insights into Theory, Methodology and Practice* (pp. 1-22). Brill.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 39-74). John Wiley & Sons.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

Tillema, M., van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2012). Quantifying the quality difference between L1 and L2 essays: A rating procedure with bilingual raters and L1 and L2 benchmark essays. *Language Testing*, *30*(1), 71–97. http://doi.org/10.1177/0265532212442647

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 & Practice*, *26*(1), 59–74. http://doi.org/10.1080/0969594X.2016.1253542

Van den Bergh, H., De Maeyer, S., van Weijen, D., & Tillema, M. (2012). Generalizability of text quality scores. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Bergh (Eds.), *Measuring Writing: Recent Insights into Theory, Methodology and Practice* (pp. 23-32). Brill.

Van Steendam, E., Tillema, M., Rijlaarsdam, G., & Van den Bergh, H. (Eds.) (2012). *Measuring Writing: Recent Insights into Theory, Methodology and Practice.* Brill.

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *26*(5), 541–562. http://doi.org/10.1080/0969594X.2019.1602027

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2017). Scale Separation Reliability: What Does It Mean in the Context of Comparative Judgment? *Applied Psychological Measurement*, *9*, 014662161774832–18. http://doi.org/10.1177/0146621617748321

Weigle, S. C. (2002). *Assessing Writing*. Cambridge University Press.

Appendix A: Writing prompt for a descriptive text

The writing prompt in Appendix A is originally developed and used by De Smedt et al., 2017.

Last week an alien landed on our planet Earth. King Filip gave him permission to visit our country to see how people live here. One day, the alien walks past a large building. He is surprised when he sees several little persons entering the building through a large gate. The alien does not know what happens in that building and what these little persons do there.

When you look at the picture, you definitely know what the alien sees. Write an informational brochure for the alien so he knows what the building is and what happens there.

Appendix B: Benchmark rating scale for descriptive writing

This benchmark rating scale is originally developed and used by De Smedt et al., 2017.

The large building that the alien sees, is a school. A school is something where you learn things, such as: math,

This large building is a school and these little

lunch. And that was everything about the school.

Thanks for reading!

A school is a large building with some kind of playground. Children from 2.5 years until 12

It's a large building

That building is a

and swim. And we eat.

70 **85** **100 (average text)** **115** **130**

Score	Strong aspects	Weak aspects
75	• *Text genre:* The writer provides information in the text. • *Assignment:* The text corresponds to the assignment; the writer describes what a school is. • *Idea development:* The ideas are factual and related to the writing topic.	• *Idea development:* The number of ideas is limited. The ideas are very general; there are no specific or remarkable ideas. • *Quality of information:* The quality of information is limited, resulting in a very vague description of what a school is. There is limited use of examples and details. • *Text structure:* The text is chaotic. It is difficult for a reader to follow the line of reasoning. • *Sentence structure and word choice:* Word choice is not varied, and the sentence structure is substandard.
85	• *Text genre:* The writer provides information in the text. • *Assignment:* The text corresponds to the assignment; the writer describes what a school is. • *Idea development:* The ideas are factual and related to the writing topic. • *Text structure:* The structure is basic. As a reader, you can follow the line of reasoning. • *Quality of information:* The text provides a minimum of information. The reader can read a very general description of what a school is.	• *Idea development:* The number of ideas is limited. The ideas are very general; there are no specific or remarkable ideas. • *Quality of information:* There is limited use of examples and details. • *Sentence structure and word choice:* Word choice and sentence structure are not varied.
100	• *Text genre:* The writer provides information in the text. • *Assignment:* The text corresponds to the assignment; the writer describes what a school is. • *Idea development:* The ideas are factual and related to the writing topic.	• *Idea development:* The number of ideas is limited. The ideas are very general; there are no specific or remarkable ideas. • *Quality of information:* There is limited use of examples and details.

	• *Text structure:* The structure is basic. As a reader, you can follow the line of reasoning. • *Quality of information:* The text provides basic information. The reader can read a general description of what a school is.	• *Sentence structure and word choice:* Word choice and sentence structure are limitedly varied.
115	• *Text genre:* The writer provides information in the text. • *Assignment:* The text corresponds to the assignment; the writer describes what a school is. • *Idea development:* The ideas are factual and related to the writing topic. At times, the writer provides remarkable ideas. • *Text structure:* The structure is basic. As a reader, you can easily follow the line of reasoning. • *Quality of information:* Next to the basic information provided in the text, the reader can read more concrete and detailed information about a school. • *Sentence structure and word choice:* The sentence structure is varied, and word choice is at times remarkable.	• *Idea development:* The number of ideas is quite limited. • *Quality of information:* The number of examples is quite limited.
130	• *Text genre:* The writer provides information in the text. • *Assignment:* The text corresponds to the assignment; the writer describes what a school is. • *Idea development:* There are a number of factual ideas that are related to the writing topic. There are also a number of remarkable ideas in the text. • *Text structure:* The text is logically structured. As a reader, you can easily follow the line of reasoning. • *Quality of information:* Next to the basic information provided in the text, the reader can read a lot of concrete and detailed	

	information about a school. The writer also provides a lot of examples. • *Sentence structure and word choice:* The sentence structure is varied, and word choice is remarkable.	