

Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner

Renske Bouwer, Monica Koster & Huub van den Bergh

To cite this article: Renske Bouwer, Monica Koster & Huub van den Bergh (2023) Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner, *Assessment in Education: Principles, Policy & Practice*, 30:3-4, 302-319, DOI: [10.1080/0969594X.2023.2241656](https://doi.org/10.1080/0969594X.2023.2241656)

To link to this article: <https://doi.org/10.1080/0969594X.2023.2241656>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 11 Aug 2023.



[Submit your article to this journal](#)



Article views: 940



[View related articles](#)



[View Crossmark data](#)



Citing articles: 2 [View citing articles](#)



Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner

Renske Bouwer , Monica Koster and Huub van den Bergh 

Institute for Language Sciences, Utrecht University, Utrecht, The Netherlands

ABSTRACT

Assessing students' writing performance is essential to adequately monitor and promote individual writing development, but it is also a challenge. The present research investigates a benchmark rating procedure for assessing texts written by upper-elementary students. In two studies we examined whether a benchmark rating procedure (1) leads to reliable and generalisable scores that converge with holistic and analytic ratings, and (2) can be used for rating texts varying in topic and genre. Results support evidence that benchmark ratings are a valid indicator of text quality as they converge with holistic and analytic scores. They are also associated with less rater variance and less task-specific variance, leading to reliable and generalisable ratings. Moreover, a benchmark scale can be used for rating different tasks with the same reliability, at least when texts are written in the same genre. Taken together, a benchmark rating procedure ensures meaningful and useful information on students' writing.

ARTICLE HISTORY

Received 30 March 2021



Accepted 24 July 2023


KEYWORDS

Writing assessment;
benchmark rating procedure;
validity; reliability;
generalisability

Assessment of writing is essential for teaching and learning to write as it provides insight in the strengths and weaknesses of students' writing performance. Teachers can use this information for formative purposes by monitoring students' writing development over time and taking informed decisions on how to further improve their writing (Black & William, 2018). The more insight teachers have in students' writing performance, the better they can adapt whole-class writing instructions and tailor feedback to students' individual needs, which ultimately leads to improvements in students' writing proficiency (Black & William, 2018; Bouwer & Koster, 2016; Graham, 2018). Therefore, it is imperative that writing assessments provide valid and reliable information about students' writing proficiency (cf. Black et al., 2011; Kane, 2016).

As writing proficiency is a construct that cannot directly be observed, information has to be inferred from the quality of students' writing (Bachman & Palmer, 1996; Weigle, 2002). However, there are two problems hampering the inferences that can be made about students' writing performance on the basis of text quality. First, there is

CONTACT Renske Bouwer  r.bouwer@uu.nl  Institute for Language Sciences, Utrecht University, Trans 10, 3512 JK Utrecht, the Netherlands

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/0969594X.2023.2241656>

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

considerable variability between raters in how they evaluate text quality, which is a threat to valid interpretations of text quality scores (i.e. scoring inference; Kane, 2013). Second, students do not perform consistently across tasks, which challenges generalisations to individual writing performance (i.e. generalisation inference; Kane, 2013). Scoring reliability and generalisability are both necessary (but not sufficient) conditions for a valid interpretation and use of text quality scores for formative purposes, see Hopster den Otter et al. (2019). That is, teachers need reliable evaluations of text quality to make an accurate diagnosis about the strengths and weaknesses in their students' writing. They also need evaluations that are representative and generalisable to select appropriate action, for instance by prioritising feedback on persistent writing problems rather than careless mistakes (Hopster den Otter et al., 2019). Thus, we must first determine how to produce consistent and generalisable scores for text quality before examining further claims about whether this information can be used to provide formative feedback to individual students. In this research, we will investigate the effects of a benchmark rating procedure on the reliability and generalisability of text quality scores, compared to holistic impressions and analytic checklists that are common in classroom practice. This comparison will provide evidence for valid interpretations within a classroom writing assessment.

Rater variability

There are two main reasons why text quality ratings vary considerably between raters. First, text quality is determined by multiple features (Huot, 1990). For instance, a text can be of high quality because ideas are well developed or structured in a logical way, but at the same time it can be poor because of errors in grammar, spelling, punctuation or conventions (or vice versa). Raters have to take all the relevant features into account when deciding whether a text meets the required standard, but they may vary in how to do so.

Second, instead of having one fixed standard of what a good text should look like, raters have their own standards of what constitutes a good text, which influence how they rate text quality and how much weight they apply to certain aspects of writing (Eckes, 2008). Such rater types were already demonstrated in 1961 by Diederich and colleagues. They asked 53 raters from different professions to holistically score the quality of 300 texts without any standards or criteria. Raters' average correlation was low, ranging from .22 for business executives to .41 for English teachers. None of the papers received less than five different grades. There were five rater types, whose primary focus varied from ideas, form, style, mechanics to the wording of a text. There is also considerable variability within raters, as reflected by individual raters who assign different scores to the same text on different occasions (Lesterhuis, 2018). It is evident that both within and between rater variability are problematic for the assessment of writing performance.

One way to reduce rater variability is to use automatic evaluations of text quality instead of human raters, for instance by software such as e-rater (Attali & Burstein, 2006). However, even though significant technological advances have been made that allow for efficient and reliable scoring of various linguistic features such as grammar, syntactic complexity, mechanics, and cohesion, these automatic measures are not comparable to

human raters as they cannot yet adequately evaluate the meaning and communicative effectiveness of students' writing (Shermis et al., 2017).

Another way to reduce rater variability is to provide raters with clear instructions and protocols on how to evaluate text quality. Such rating procedures are considered to reduce rater variability by limiting the freedom of the rater during the rating process (Wesdorp, 1981). When this is preceded by a training in which raters can collectively rate, compare and discuss various example texts using the predefined criteria or scale points rater variability may be further reduced (Weigle, 1994). But because of the considerable time investment that this requires, such extensive training sessions are hardly feasible in classroom practice (Inspectorate of Education, 2021). Also, training does not guarantee that raters interpret and apply scoring criteria in the same manner, as trained raters may still differ in their focus and decision-making processes (Eckes, 2008). It is therefore necessary to compare different rating procedures that may provide feasible but effective ways to support agreement between untrained raters.

Holistic and analytic rating procedure

The two most common rating procedures in classroom practice are holistic and analytic ratings. In a holistic rating procedure, teachers are instructed to assign text quality scores based on their general impression of the text-as-a-whole, often with predefined rating criteria but without having to score criteria separately (Huot, 1990). An example is primary trait scoring, in which raters only have to evaluate whether the writer successfully accomplished the rhetorical purpose of the writing task (Lloyd-Jones, 1977). In an analytic rating procedure, all aspects of text quality are rated separately (e.g. content, structure, style or mechanics), which are then combined in an overall text quality score. Analytic rating procedures can also prescribe how much weight raters should give to each aspect, such that higher level aspects as content and structure receive relatively more weight than lower-level aspects like grammar and spelling (cf. Barkaoui, 2011). Even though researchers have questioned whether text quality can be reflected by the sum of its parts (Huot, 1990; Sadler, 2009), analytic scores can provide fine-grained insight in the specific strengths and weaknesses of individual writing, which is important to support learning in the classroom (Hopster den Otter et al., 2019).

There are different kinds of analytic scales, such as criteria-lists, rubrics or checklists, which vary in the task-specificity and restrictiveness of scoring instructions (Jönsson & Svingby, 2007; Weigle, 2002). For instance, analytic rubrics are more restrictive for raters than rating text quality based on holistic impressions as they prescribe both the criteria that need to be scored and the possible performance levels for each criterion. Because of these rater restrictions, rubrics are supposed to reduce rater variability to a larger extent. However, previous findings have been inconsistent, suggesting that analytic ratings are no guarantee for high inter-rater reliability (e.g. Barkaoui, 2011). For instance, in the context of large-scale writing assessments, thinking-aloud protocols revealed that raters vary in how they interpret and use the criteria and ratings scales in analytic rubrics (Lumley, 2002). Even with such specific rubrics, raters tend to use their holistic impression of text quality to rate its separate features. This was especially the case for higher order criteria, such as ideas or structure in a text. This does not only explain why rater

variance remains for analytic rubrics, but also challenges the assumption that text features can be considered as independent constructs at all, and hence can be scored analytically. These results also emphasise the need for training to obtain reliable ratings with analytic rubrics (Grabowski et al., 2014; Jönsson et al., 2021).

Analytic checklists are developed with the aim to restrict the freedom of raters even further. This is particularly relevant for classroom contexts and high-stake tests in which it is not feasible to provide extensive training sessions for individual raters. In analytic checklists, raters only have to indicate whether certain features are present in the written text. Such simple dichotomous judgements appear to improve the consistency of text quality ratings without the need for extensive training (Grabowski et al., 2014). Uzun et al. (2019) also showed that checklists reduce rater variability in comparison to analytic rubrics. However, a downside of such a strict rating method is the large task-specificity, which challenges the generalisations that can be made to individual writing proficiency beyond the writing task.

Previous generalisability studies have indeed shown that it is hard to make generalisations about students' writing proficiency on the basis of one written text, rated by one rater (e.g. Barkaoui, 2007; Bouwer et al., 2015; Schoonen, 2005; Van den Bergh et al., 2012). This is the case for analytic ratings due to large variance between tasks, as well as for holistic ratings in which rater variance is often larger than the variance due to the writer. This may limit valid and reliable interpretations of both holistic impressions and analytic checklists.

Benchmark rating procedure

An alternative way to reduce rater variance is to implement a benchmark rating procedure. In this comparative approach, text quality is rated by comparing each text to benchmark texts on a continuous rating scale (Schoonen, 2005; Wesdorp, 1981). If carefully selected, these benchmarks are representative for the range of text quality that is produced in a particular age-group, and hence, are illustrative for levels of low, average, and high text quality. By comparing students' texts to these benchmarks, instead of (only) to each other, raters make a constant comparison to a predetermined criterion, which allows for criterion-referenced scoring (cf. Marshall et al., 2020). Even though raters are provided with a standard of increasing levels of text quality in which analytic features are operationalised, explained and illustrated, a benchmark rating procedure facilitates holistic evaluations, and might therefore be considered to be 'the best of both worlds'. It may also be considered a more natural way of rating text quality, as every evaluative judgement is actually a comparison, either to an internal standard, or to previously seen work (Laming, 2004; Pollitt, 2012; Sadler, 1998).

Benchmark rating procedures have already been successfully used in previous research to evaluate students' writing performance with sufficient reliability (Bouwer et al., 2018; De Smedt et al., 2016; Rietdijk et al., 2017; Tillema et al., 2012). Benchmarks were also successfully applied in large-scale assessments such as the National Assessment of Educational Progress (NAEP, e.g. Solomon et al., 2004), to support raters to reliably rate holistic text quality. Yet, there is hardly any research in which holistic benchmark ratings are compared to rating procedures that are more common in classroom practice, such as ratings based on

holistic impressions or analytic checklist without any benchmarks. In particular, it is yet unknown whether benchmark ratings measure text quality in a similar way (i.e. convergent evidence) as holistic and analytic ratings, but with higher levels of reliability and generalisability, and hence can be validly interpreted as information about students' writing proficiency.

In addition, it is still unknown whether a benchmark rating scale can be used to rate the quality of different kinds of texts. Several researchers have hypothesised that a benchmark rating procedure can only support raters when the to-be-rated texts are similar to the benchmarks (Feenstra, 2014; Pollman et al., 2012). This would imply that for each writing task a unique benchmark scale has to be developed, which takes time and expertise and strongly impedes large-scale use of benchmark rating scales in practice (Feenstra, 2014; Pollman et al., 2012). However, up until now, this hypothesis has not been empirically tested.

Research aim

The twofold research aim is to examine whether a benchmark rating procedure (1) leads to reliable text quality scores that converge with holistic and analytic ratings and can be generalised across tasks, and hence, can be validly interpreted as students' writing proficiency, and (2) can be used for rating text quality for tasks that differ in topic and genre from the benchmarks. These research questions were investigated in two separate studies. For study 1, we hypothesised that benchmark ratings are more reliable and generalisable than holistic and analytic ratings, as raters are expected to experience sufficient support by comparing texts to benchmarks instead of forming a general impression of text quality in an absolute manner, and without having to break text quality down in analytical parts. For this reason, we also expected that benchmark ratings highly converge with holistic ratings. Finally, we expected that the similarity in topic and genre between benchmarks and to-be-rated texts has an effect on the reliability of benchmark ratings.

Study 1

In study 1, we compared text quality scores obtained via a benchmark rating procedure to those of a holistic and analytic rating procedure, in order to examine its reliability, convergent validity, and generalisability.

Participants

A total of 36 undergraduate students (31 female) from the Department of Language, Literature and Communication participated voluntarily as raters in this study. Their mean age was 22.62 years ($SD = 3.41$). Raters were randomly assigned to one of three rating procedures: holistic ($n = 11$), benchmark ($n = 12$) or analytic ($n = 13$).

Materials and procedure

Text samples

From the total of 99 upper primary students who participated in a previous research project (see Pullens, 2012), we randomly selected 34 students who completed two persuasive writing tasks. This resulted in two samples of 34 persuasive texts of low, average as well as high quality. In both writing tasks, students had to write a formal letter to a fictional company about a problem with a promotion campaign. One task was about collecting points on Yummy Yummy candy bars (Yummy, see Appendix A) and the other task was about collecting Smurfs in the supermarket (Smurf, see Appendix B). The handwritten texts of the students were retyped including all errors concerning spelling, grammar, punctuation or capitals made by the student. Raters received either the Yummy or the Smurf text sample and they were instructed to individually rate the quality of the texts in their sample using the assigned rating procedure. In all three conditions, raters had to start rating immediately, without any training. This reflects current practice, in which teachers hardly discuss ratings or exemplars with colleagues to practice using a rating scale (cf. Inspectorate of Education, 2021). They had to write down their scores on the text and they were allowed to change scores only in exceptional cases. The rating process took about one hour.

Holistic rating procedure

Participants assigned to the holistic rating procedure were instructed to provide a general impression of text quality on a scale from 1 to 10, with the following criteria in mind: (1) content: a clear problem statement and question to the reader, (2) structure: a coherent and understandable text, (3) style: effective and varied language use, appropriate for the intended audience and purpose, (4) genre conventions: meeting the formal requirements (e.g. contact details, date and proper salutation and closure), (5) mechanical aspects: sufficient use of grammar, spelling, and punctuation rules. Raters were free in how they weighted these features into an overall score for text quality. This holistic rating procedure and the 10-point rating scale reflects common writing assessment practice in Dutch primary schools (Inspectorate of Education, 2021).

Analytic rating procedure

Participants assigned to the analytic rating procedure received a checklist in which they had to indicate the presence or absence of 15 specific text features, see Appendix D. This checklist was developed and used by the Dutch Institute of Educational Measurement to evaluate primary students' level of writing proficiency in a large-scale national assessment (Kuhlemeier et al., 2013). There were 6 items on the content of the text (e.g. does the student state something about the eight points already collected?), 6 items were on structure and conventions (e.g. is there a formal salutation?), and 3 items were related to the communicative goal and audience (e.g. are the arguments convincing?). Raters had to score each item with one point if the particular feature was present in the text and with zero points if it was absent. The three dimensions strongly correlate, indicating that they can be combined into one overall score for text quality (Kuhlemeier et al., 2013). The total score for text quality was based on the aggregation of the scores for the 15 items in the scoring form, resulting in overall text quality scores ranging from 0 to 15.

Benchmark rating procedure

Participants assigned to the benchmark rating procedure received a continuous interval rating scale ranging from 0 to infinite, with an arbitrary mean of 100, a standard deviation of 15, and five benchmarks that represent the increasing levels of text quality for students in grade 4 to 6. The benchmark of average quality marked the centre position on the rating scale (i.e. 100 points). The other benchmarks were one (115 points) and two (130 points) standard deviations above average, and one (85 points) and two (70 points) standard deviations below average. For each benchmark was described why the text was representative for its location on the scale, using the same criteria as the holistic rating procedure. Raters had to score each student text holistically by comparing it to the benchmarks and place it on the interval rating scale accordingly. They were allowed to provide all possible scores, either within or outside the range of the benchmarks. However, assuming a normal distribution of text quality scores, extreme low or high scores are very unlikely.

There were two benchmark rating scales in this study, see Appendix C for the benchmark rating scale of the Yummy task. The benchmarks originated from the same project as the text samples, and they were selected on the basis of scores provided by juries of three independent raters ($\rho = .77$; see Pullens, 2012). To ensure equal intervals between the benchmarks on the rating scale, we used two selection criteria. First, we used the normal distribution of the text quality scores to select several texts at -2 SD, -1 SD, 0, $+1$ SD, $+2$ SD. Second, we selected for each performance level the benchmark text that was most representative and with high agreement between the three jury-raters (i.e. low variance between raters' scores).

Data analysis

Inter-rater reliability was estimated by Cronbach's alpha, indicating the consistency of ratings from independent raters. We considered coefficient values of 0.7 as acceptable and values of 0.8 or higher as an indicator of good rater agreement (cf. Stemler, 2004). Because Cronbach's alpha increases by the number of raters, we used the Spearman-Brown formula¹ to determine the reliability coefficients for an equal number of raters for each rating procedure. Further, we applied a K-sample significance test to compare the reliability coefficients of text scores for the holistic, analytic and benchmark rating procedure (Feldt, 1980; Hakstian & Whalen, 1976).

Convergent validity evidence is obtained by measuring the degree to which text quality scores of the holistic, analytic and benchmark rating procedure are correlated. High correlation coefficients between rating procedures indicate that they measure the same construct (Cook & Campbell, 1979). As all three rating procedures in the present study are developed to measure the construct of text quality, it is assumed that they will correlate highly with each other. Low correlations impose a threat to construct validity, indicating that at least one of the rating procedures measures partly a different construct. Since the rating procedures will not have a perfect reliability due to measurement error, correlations between scores from two rating procedures will likely suffer from attenuation. Therefore, we corrected for attenuation attributable to unreliability by dividing the observed correlation coefficient by the product of the square roots of the two relevant

Cronbach's alpha reliability coefficients (Lord & Novick, 1968). These attenuated correlation coefficients reflect the true correlations between rating procedures.

Generalisability theory is used to estimate the generalisability of text quality scores for each rating procedure (Brennan, 2001; Cronbach et al., 1972). First, we disentangled different sources of variation in the measurement of writing. In this study, the sources of variation included the writer, task, writer-by-task interaction, rater-by-task interaction, and three-way interaction between writer, task and rater, including random error. Because raters were nested within writing tasks, the estimation of rater effects is contaminated by rater-by-task interaction. Estimated variance components were computed for each source of variation using SPSS, separately for each rating procedure. Second, we approximated the generalisability coefficient for each rating procedure, by estimating the variance that is associated with the writer (i.e. true score) as a proportion of the total amount of variance (including all sources that are regarded as measurement error). Hence, generalisability coefficients indicate the extent to which text quality scores can be generalised to students' writing proficiency. Third, we compared the generalisability coefficients, in order to determine whether the generalisability of writing scores depends on the rating procedure.

Results and conclusions

Reliability

Means and standard deviations, as well as the inter-rater reliability coefficient for holistic, analytic, and benchmark ratings are presented in Table 1. Average scores for the Smurf task were somewhat lower than the average scores for the Yummy task. Results also show that Cronbach's alpha reliability coefficient ranged from .79 to .93, depending on the rating procedure and writing task.

As the number of raters for each rating procedure was not equal, we used the Spearman-Brown formula to estimate reliability coefficients for scores based on one, two and three raters, see Table 2. Results showed that for none of the rating procedures a desired reliability level of .70 was reached when only one rater is involved in the

Table 1. Means, SD and Cronbach's Alpha Reliability for Holistic, Benchmark, and Analytic Ratings.

Rating procedure	Scale	Task Yummy				Task Smurf			
		<i>N</i>	Mean	<i>SD</i>	α	<i>N</i>	Mean	<i>SD</i>	α
Holistic	0–10	4	5.93	1.67	.79	7	4.94	1.86	.91
Benchmark	0 - ∞	7	91.00	20.48	.93	5	82.59	19.88	.82
Analytic	0–15	7	8.76	3.01	.93	6	8.30	2.72	.87

Table 2. Average Reliability Coefficients for One, Two and Three Raters.

Rating procedure	Yummy			Smurf		
	Reliability (ρ)			Reliability (ρ)		
	1 rater	2 raters	3 raters	1 rater	2 raters	3 raters
Holistic scale	.48	.65	.74	.60	.75	.82
Benchmark scale	.64	.78	.84	.48	.65	.74
Analytic scale	.64	.78	.84	.54	.70	.78

Table 3. Correlations between Holistic, Benchmark, and Analytic Ratings.

	Yummy			Smurf		
	Holistic	Benchmark	Analytic	Holistic	Benchmark	Analytic
Holistic	–	.90	.83	–	.83	.63
Benchmark	1.00	–	.86	.96	–	.62
Analytic	.97	.93	–	.86	.73	–

Note. Uncorrected correlations are above the diagonal and attenuated correlations are below the diagonal.

measurement, which is often the case in educational practice. Also, there were no significant differences between rating procedures. For the Yummy task, the reliability of benchmark ratings based on a single rater was .64, which was not significantly higher than the reliability of holistic ratings (.48, $F(1, 33) = 1.44, p = .15$), or analytic ratings (both .64, $F(1, 33) = 1.00, p = .50$). For the Smurf task, the reliability of the benchmark ratings was .48, which also did not differ significantly from holistic ratings (.60, $F(1, 33) = 1.30, p = .23$) or analytic ratings (.48 and .54, $F(1, 33) = 1.13, p = .36$). Table 2 also shows that regardless of the rating procedure, at least two raters are required to reach a sufficient level of reliability of .70. For a reliability of .80 or higher, ratings should be based upon judgements of three or more raters.

Convergent validity

Table 3 shows that the (attenuated) correlations between the three rating procedures were high. This indicates that benchmark ratings are converging to the same construct as holistic and analytic ratings, which provides evidence for a valid interpretation of benchmark ratings as a measure of text quality. There were also some differences. For Yummy, for instance, the average attenuated correlation between holistic, benchmark and analytic ratings was .98 and for Smurf .79. Further, the correlations between benchmark and holistic scores were higher than the correlations between benchmark and analytic scores. This was even stronger for the Smurf task, compared to the Yummy task. This may indicate that the benchmark rating procedure allows raters to make holistic comparisons, and hence evaluate the quality of texts as a whole, rather than making analytical comparisons using quality criteria.

Generalizability

Table 4 shows for each rating procedure the proportion of variance that is associated with writer, task, writer-by-task-interaction, rater-by-task-interaction, and three-way interaction including random error. Results show that the percentage of variance that is related to the writer was highest for the benchmark rating procedure, which makes it easier to

Table 4. Variance Components as Proportions of the Total Variance for each Rating Procedure.

Source	Holistic	Benchmarks	Analytic
Student (s)	.23	.33	.28
Task (t)	.06	.06	.00
Student by task (st)	.10	.18	.27
Rater within task (r:t)	.35	.08	.08
Student by rater within tasks, and error (s(r:t),e)	.26	.35	.37

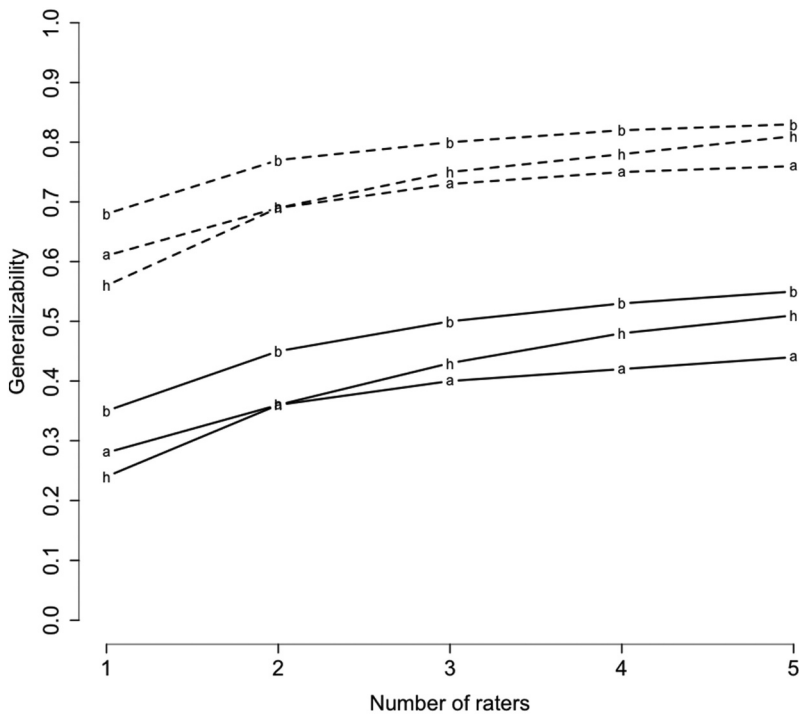


Figure 1. Estimated generalisability of writing scores for one task (solid lines) and four tasks (dashed lines), with varying number of raters. Lines represent the rating procedure that is used by raters, b: benchmark rating procedure, a: analytic rating procedure, h: holistic rating procedure.

generalise to students' overall writing performance. More particularly, for benchmark ratings, 33% of the variance in text quality scores can be explained by differences between writers, whereas the writer-related variance for holistic and analytic ratings was respectively 23% and 28%. Hence, if one random rater uses the benchmark procedure to rate two texts written by the same writer, the expected correlation between will equal $\sqrt{.33}$, which is .57. The expected correlation between two texts written by the same writer will be lower when the rater evaluates holistically ($\sqrt{.23} = .48$) or analytically ($\sqrt{.28} = .53$).

The variance components in Table 4 further show that holistic ratings included large rater-within-task variance: 35% versus 8% for benchmark and analytic ratings. This indicates that a holistic rating procedure yields relatively rater-specific judgements, which makes generalisation across raters rather difficult. It was also demonstrated that analytic ratings included relatively large student-by-task interaction: 27% versus 10% and 18% for holistic and benchmark ratings respectively. This indicates that an analytic rating procedure yields relatively task-specific judgements, which makes generalisation across tasks more difficult.

Based upon these variance components, we approximated how many writing tasks and raters are needed to make generalisations about students' writing performance. For this, we estimated the generalisability coefficient, which reflects the proportion of writer variance to the sum of all variances. Figure 1 shows the generalisability coefficients for the three rating procedures and a varying number of tasks and raters.

It shows that although benchmark ratings are more generalisable than analytic or holistic ratings, a desired level of generalisability of at least .70 is not reached when the assessment includes only one writing task. The generalisability increases when more tasks are included. For a generalisability of at least .70, students have to write at least four tasks, rated by two different raters using the benchmark rating procedure. More raters (at least three) are needed when text quality is rated in a holistic or analytic manner.

Study 2

In this study we examined whether the same benchmark scale can be used for rating different kinds of texts. Therefore, we compared the inter-rater reliability of text quality ratings for texts that (a) were similar to the benchmarks on the scale regarding both topic and genre, (b) differed in topic, but were the same genre as the benchmarks, and (c) differed in both topic and genre.

Participants

Ten undergraduate students (9 female, 1 male) from the Department of Language, Literature and Communication participated voluntarily as raters in this study. Their mean age was 21.90 years ($SD = 1.73$). None of the participants indicated to have experience with rating the quality of texts written by students in elementary grades.

Materials and procedure

Raters rated the quality of three text samples that each consisted of 40 texts written by upper-elementary students, i.e. 120 texts in total. These texts were randomly selected from three writing tasks that varied in topic and/or genre, from earlier research projects on upper-elementary students' writing development (Bouwer & Koster, 2016; Pullens, 2012). For two writing tasks, students had to write a persuasive letter to a fictional company about a problem with a promotion campaign. These writing tasks were similar with regard to genre and differed only in topic. One task (Yummy) was about collecting Yummy Yummy candy bars for a music CD. The other task (Smurf) was about collecting Smurfs for a digital camera. Task descriptions are included in Appendices A and B. The third writing task (Like) differed from the other tasks in topic as well as in genre: students had to write a letter of advice to a fictitious peer on how to get good grades for writing. Description of this task is included in Appendix E.

Raters used a similar benchmark rating procedure as in Study 1 to rate the quality of 120 student texts. Two different benchmark rating scales were used in this study, one with benchmarks from the Yummy task and one with benchmarks from the Smurf task, see Appendix C for an example. Half of the raters ($n = 5$) received the Yummy benchmark scale and the other half ($n = 5$) received the Smurf benchmark scale. As in Study 1, the benchmark rating scale is considered as a continuous interval scale. Thus, even though the five benchmarks mark different quality levels on the rating scale (i.e. 70, 85, 100, 115 or 130 points), raters were allowed to provide all possible scores ranging from 0 to infinite, either within or outside the benchmark levels.

Raters rated the texts in a specified order. First, they rated texts that were similar to the benchmarks in both topic and genre. Second, they rated texts that were similar to the

Table 5. Means, SD and Reliability by Benchmark Condition and Text Sample.

Similarity	Yummy benchmark scale				Smurf benchmark scale			
	<i>N</i>	Mean	<i>SD</i>	<i>α</i>	<i>N</i>	Mean	<i>SD</i>	<i>α</i>
Topic + genre	5	89.87	18.18	.92	5	94.09	14.53	.91
Genre	5	90.50	17.44	.90	5	88.85	17.81	.92
None	5	90.83	15.95	.87	5	88.43	15.97	.84

benchmarks in genre but differed in topic. Third, they rated texts that were different in both genre and topic. Thus, raters with the Yummy benchmark scale first rated the Yummy texts, then the Smurf texts and finally the Like texts, while raters with the Smurf benchmark scale first rated the Smurf texts. The order of texts within samples was kept the same for all raters.

All raters received written instructions and a short training of 15 minutes before starting with rating the quality of the texts. The Yummy and Smurf groups were trained separately. Training involved discussion of the benchmarks and guidelines of the rating procedure. Raters were instructed to work independently, and to focus primarily on the communicative effectiveness of the text while comparing it to the benchmarks on the rating scale. They practiced scoring of three selected example texts, one of poor quality, one of average quality and one of high quality. The example texts were from the same task as the benchmarks on the scale. Further, they were instructed to rate texts one by one, and they were not allowed to adjust their ratings once they were given. They had a maximum of 120 minutes to rate all texts and were allowed to take individual breaks. After the ratings, participants were asked to indicate on a 5-point Likert scale how they perceived the difficulty of the rating procedure for the three different samples of texts.

Data analysis

Inter-rater reliability for the benchmark ratings were estimated per rater group and text sample, using Cronbach's alpha. A K-sample significance test was used to compare the Cronbach's alpha reliability coefficients between the three text samples (Feldt, 1980; Hakstian & Whalen, 1976).

Results and conclusions

Means, standard deviations and inter-rater reliability coefficients for the three text samples and two rater groups are presented in Table 5. There were no significant differences in the inter-rater reliability for texts that were similar to the benchmarks in both topic and genre versus texts that were only similar in genre, with a Cronbach's alpha of respectively .92 versus .90 for the Yummy benchmark scale ($F(1, 39) = 1.21, p = .27$), and .91 versus .92 for the Smurf benchmark scale ($F(1, 39) = 1.23, p = .26$). However, raters' scores were less consistent for texts that were from a different genre than the benchmarks on the scale, at least for the raters who used the Smurf benchmark rating scale. More particularly, for these raters, results show that the inter-rater reliability for texts of a different genre was significantly lower than the reliability of texts in the same genre as the benchmarks (Cronbach's alpha of .84 versus .92, $F(1, 79) = 1.81, p < .01$). Although the reliability for raters using the Yummy benchmark rating scale was also

relatively low when texts of a different genre were rated in comparison to texts of the same genre (respectively $\alpha \geq .87$ and $.91$), this difference failed to reach significance ($F(1, 39) = 1.52, p = .10$).

These results were supported by the questionnaire data, which revealed that raters experienced more difficulties when they had to rate texts in another genre ($M = 1.30, SD = 0.21$), than when they had to rate texts that were similar to the benchmarks only in genre ($M = 3.50, SD = 0.21$) or in both genre and topic ($M = 3.90, SD = 0.21$), $F(2, 27) = 46.02, p < .001$.

General discussion

The present research examined, in two separate studies, the extent to which a benchmark rating procedure supports raters in providing text quality scores that converge with holistic and analytic text quality ratings and are both reliable and generalisable, also for rating different kinds of texts. Study 1 showed high correlations between benchmark ratings and ratings based on holistic impressions or analytic checklists, which provides convergent evidence that benchmark ratings can be validly interpreted as measures of text quality. Furthermore, it was shown that inter-rater reliability for benchmark ratings was high, but only for average scores of at least two raters. Similar results were found for holistic and analytic ratings, indicating that regardless of the rating procedure, text quality ratings based on only one rater are insufficiently reliable. Moreover, the generalisability study revealed that in addition to multiple raters, multiple tasks are required for reliable and generalisable results. There was, however, an effect of rating procedure on the number of raters and tasks needed for a sufficient level of generalisability. In particular, benchmark ratings were associated with less rater variance compared to ratings based on holistic impressions as well as with less task-specific variance than ratings based on analytical checklist. As a result, a benchmark rating procedure requires fewer raters and tasks than a holistic or analytic rating procedure to obtain comparable levels of generalisability of text quality scores. Study 2 provided evidence for the practical use of the benchmark rating procedure, as high levels of inter-rater reliability were also obtained for texts that were different from the benchmark on the scale. However, benchmarks were slightly less supportive to raters when they had to rate texts in a different genre.

Taken together, the two studies provide evidence that a benchmark rating procedure ensures meaningful and interpretable information on the quality of student's writing, which is both reliable and generalisable, and hence can be used for making inferences about student's writing proficiency across tasks and raters. This forms the basis for well-justified summative decisions, as well as effective classroom instruction and individual feedback (Black & William, 2018; Hopster den Otter et al., 2019). However, the present research showed that, even for benchmark ratings, at least four tasks and two raters are needed to allow for generalisations about students' writing proficiency. This confirms previous research, showing consistently that multiple tasks and raters are needed in the assessment of writing (Bouwer et al., 2015; Schoonen, 2005; Van den Bergh et al., 2012). This has important implications for educational practice in which students' writing performance often is assessed by only one teacher and one task.

The present research also provided evidence regarding the practical use of benchmark ratings. First, benchmark ratings were associated with a higher generalisability than ratings based on holistic impressions or analytic checklists. This implies that fewer tasks and raters are needed for valid and reliable decisions about students' level of writing. Second, the same benchmark scales can be used for rating different writing tasks, at least when texts are written in the same genre. This makes it more feasible to include multiple tasks in the writing assessment. Although in this study the order in which texts were rated was confounded with the similarity between the texts and benchmarks (i.e. raters always rated more similar texts before they rated different texts, which might have affected the outcomes), we were able to replicate the findings for two different benchmark scales. This strongly suggests that previously raised concerns on the limited use of benchmark scales are not warranted (Feenstra, 2014; Pollman et al., 2012), and that a benchmark scale can be considered as a steady reference framework that provides clear standards for evaluating different kinds of texts.

That a benchmark scale, once constructed, can be used for more than one task is a huge benefit for application in educational practice. Benchmark selection is a demanding process, which requires time and expertise, in which teachers now only have to invest once for each genre. However, constructing a benchmark rating scale is difficult to implement in educational practice, for detailed instructions see Feenstra (2014) or Pollman et al. (2012). This is especially true when we bear in mind that assessors have to agree on the quality of the examples in the benchmark scale, which becomes even more cumbersome as the number of benchmarks increases. What might be a useful procedure in classroom practice is to start with one benchmark of average quality, and gradually increase this 'scale' with more benchmarks. Another possibility is to derive writing assignments and benchmarks from large-scale national or international assessments, such as the National Assessment of Educational Progress (NAEP, e.g. Solomon et al., 2004). The benefit of these large-scale assessments is that both the writing assignments and the rating procedures have been tested rigorously and that the benchmarks are already standardised (cf. Smith & Paige, 2019).

Another advantage of the implementation of benchmark scales in educational practice is that this rating procedure enables teachers to evaluate and monitor their students' writing progress over time or to provide tailored feedback to individual students by illustrating the gap between current and desired writing performance (Sadler, 1998). Benchmarks can also be used as scaffolds for students to improve their text, or to write a better text next time. The results from the present study are therefore promising for large-scale implementation of benchmark scales for formative purposes. In a recent project with secondary school teachers, we have tested the use of benchmark ratings in the classroom (Bouwer & Gerits, 2022). Results of this project showed that teachers were positive about this rating procedure; it supported them to make reliable judgements with more certainty and in less time than the more commonly used holistic ratings or analytic rubrics. The benchmarks also provided students with specific insight in what they were already doing well and how they could improve their text. In line with study 2, the teachers reported that they were able to use the same benchmark scale for evaluating different types of texts, allowing them to monitor and compare students' writing performances over time and across classes. Further research

is needed to establish the degree to which benchmark scales can be validly used by teachers and students for formative purposes, and hence, improve students' writing performance over time.

One of the limitations of benchmark ratings is that benchmarks, even when they are carefully selected, do not automatically provide information on how students perform in relation to writing standards. For instance, an average score on a benchmark scale does not necessarily indicate that a student is an average writer. This requires a careful standardisation of the benchmarks, which is often the case in large-scale assessments such as NAEP (Solomon et al., 2004). Further research on how to standardise and normalise benchmark rating scales in educational practice is needed, in order to also enable decisions on whether students meet the desired performance levels (cf. Jönsson & Svingby, 2007).

A second limitation is that a benchmark rating procedure does not completely solve the rater problem (Barkaoui, 2011; Diederich et al., 1961; Weigle, 2002). There were still considerable differences between raters, implying that at least two raters have to be involved for a satisfactory level of reliability. In addition, we used novice raters instead of experienced teachers in this study, which limits the generalisations of the conclusions. However, teachers cannot always be considered as experienced raters. In fact, teachers report to lack sufficient training in the assessment of writing (Inspectorate of Education, 2021). It is therefore promising that, in the present study, novice raters were able to rate text quality in a reliable manner, even without training sessions. Based on this, we can assume that benchmarks will also provide adequate support to teachers who are not specifically experienced in rating text quality. Further research should establish the reliability of text quality scores when different points on a holistic, analytic and benchmark rating scale are explained, practiced, and discussed in a training session using exemplars. This will enhance a common interpretation of how to reliably interpret and apply the levels of a rating scale, both with and without rater training.

To conclude, Cooper and Odell stated already in 1977, that 'since writing is an expressive human activity, we believe the best response to it is areceptive, sympathetic human response' (p. xii). The present study contributes to making the human response more reliable and generalisable by supporting raters with benchmarks that illustrate different writing performance levels, which they can use to rate different kinds of texts. By doing so, a benchmark rating procedure leads to text quality scores that are a valid indication of students' writing performance and can be used in educational practice for summative as well as formative purposes.

Note

1. When k is the factor by which the length of the test is changed, and r_x is the reliability of the original test, the reliability of r_{kx} can be estimated by: $r_{kx} = (k * r) / (1 + (k-1) * r)$

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Netherlands Organization for Scientific Research [411-11-859].

Notes on contributors

Dr. Renske Bouwer is Assistant Professor in Language and Education at Utrecht University. She combines theories from educational sciences, psychology, and linguistics to improve the writing skills of students in all educational levels. She also investigates the merits of comparative assessment methods for the formative and summative assessment of writing.

Dr. Monica Koster is a researcher, educational developer, and teacher trainer in the field of language education, and writing skills in particular.

Dr. Huub van den Bergh is Full Professor in pedagogy measurement of language skills at the Faculty of Humanities at Utrecht University. His research focusses on methodology and statistics on the one hand and learning processes on the other.

ORCID

Renske Bouwer  <http://orcid.org/0000-0003-0434-0224>

Huub van den Bergh  <http://orcid.org/0000-0002-1320-5334>

References

- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment*, 4(3), 1–30.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.
- Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, 12(2), 86–107. <https://doi.org/10.1016/j.asw.2007.07.001>
- 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. <https://doi.org/10.1080/0969594X.2010.526585>
- 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(4), 451–469. <https://doi.org/10.1080/0969594X.2011.557020>
- Black, P., & William, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, 25(6), 551–575. <https://doi.org/10.1080/0969594X.2018.1441807>
- Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, 32(1), 83–100. <https://doi.org/10.1177/0265532214542994>
- Bouwer, R., & Gerits, H. (2022). Aan de slag met het schoolexamen schrijfvaardigheid [Getting started with the school exams for writing skills]. *Levende Talen Magazine*, 109(2), 10–15.
- 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. <https://doi.org/10.1037/edu0000206>
- Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag. <https://doi.org/10.1007/978-1-4757-3456-0>

- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Houghton Mifflin.
- Cooper, C. R., & Odell, L. (1977). *Evaluating writing: Describing, measuring, judging*. National Council of Teachers of English.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. John Wiley.
- De Smedt, F., Van Keer, H., & Merchie, E. Student, teacher and class-level correlates of Flemish late elementary school children's writing performance. (2016). *Reading & Writing*, 29(5), 833–868. <https://doi.org/10.1007/s11145-015-9590-z>
- Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability*. (Research Bulletin RB-61-15). Educational Testing Service.
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155–185. <https://doi.org/10.1177/0265532207086780>
- Feenstra, H. (2014). *Assessing writing ability in primary education. On the evaluation of text quality and text complexity* [Unpublished doctoral dissertation]. University of Twente.
- Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika*, 45(1), 99–105. <https://doi.org/10.1007/BF02293600>
- Grabowski, J., Becker-Mrotzek, M., Knopp, M., Jost, J., & Weinzierl, C. (2014). Comparing and combining different approaches to the assessment of text quality. In D. Knorr, C. Heine, & J. Engberg (Eds.), *Methods in writing process research* (pp. 147–165). Lang.
- Graham, S. (2018). Instructional feedback in writing. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 145–168). Cambridge University Press. <https://doi.org/10.1017/9781316832134.009>
- Hakstian, A. R., & Whalen, T. E. (1976). A K-sample significance test for independent alpha coefficients. *Psychometrika*, 41(2), 219–231. <https://doi.org/10.1007/BF02291840>
- Hopster den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2019). A general framework for the validation of embedded formative assessment. *Journal of Educational Measurement*, 56(4), 715–732. <https://doi.org/10.1111/jedm.12234>
- Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, 60(2), 237–263. <https://doi.org/10.3102/00346543060002237>
- Inspectorate of Education. (2021). *Peil.Schrijfvaardigheid: Einde (speciaal) basisonderwijs 2018-2019* [Level.Writing ability: End of (special) elementary education 2018-2019].
- Jönsson, A., Balan, A., & Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assessment in Education: Principles, Policy & Practice*, 28(3), 212–227. <https://doi.org/10.1080/0969594X.2021.1884041>
- Jönsson, 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>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23(2), 198–211. <https://doi.org/10.1080/0969594X.2015.1060192>
- Kuhlemeier, H., Til, A. V., Hemker, B., de Klijin, W., & Feenstra, H. (2013). Balans van de schrijfvaardigheid in het basis- en speciaal basisonderwijs 2. Periodieke Peiling van het Onderwijsniveau (No. 53). [Present state of writing competency in elementary and special education 2. Periodical assessment of the level of education]. Cito.
- Laming, D. R. J. (2004). *Human Judgment: The Eye of the Beholder*. Thomson Learning.
- Lesterhuis, M. (2018). *The validity of comparative judgement for assessing text quality: an assessor's perspective* [Unpublished doctoral dissertation]. University of Antwerp.
- Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 33–68). National Council of Teachers of English.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

- Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19(3), 246–276. <https://doi.org/10.1191/0265532202lt230oa>
- Marshall, N., Shaw, K., Hunter, J., & Jones, I. (2020). Assessment by comparative judgement: An application to secondary statistics and English in New Zealand. *New Zealand Journal of Educational Studies*, 55(1), 49–71. <https://doi.org/10.1007/s40841-020-00163-3>
- 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>
- Pollman, E., Prenger, J., & De Glopper, K. (2012). Het beoordelen van leerlingteksten met behulp van een schaalmodel [Rating student' texts with a benchmark scale]. *Levende Talen Tijdschrift*, 13(3), 15–24.
- Pullens, T. (2012). *Bij wijze van schrijven: Effecten van computerondersteund schrijven in het primair onderwijs* [In a manner of writing: Effects of computer-supported writing in primary education] (Unpublished doctoral dissertation). Utrecht University.
- Rietdijk, S., Janssen, T., van Weijen, D., van den Bergh, H., & Rijlaarsdam, G. (2017). Improving writing in primary schools through a comprehensive writing program. *Journal of Writing Research*, 9(2), 173–225. <https://doi.org/10.17239/jowr-2017.09.02.04>
- Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, 5(1), 77–84. <https://doi.org/10.1080/0969595980050104>
- Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, 34(2), 159–179. <https://doi.org/10.1080/02602930801956059>
- Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, 22(1), 1–30. <https://doi.org/10.1191/0265532205lt295oa>
- Shermis, M. D., Burstein, J., Elliot, N., Miel, S., & Foltz, P. W. (2017). Automated writing evaluation: An expanding body of knowledge. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 395–410). The Guilford Press.
- Smith, G. S., & Paige, D. D. (2019). A study of reliability across multiple raters when using the NAEP and MDFS rubrics to measure oral reading fluency. *Reading Psychology*, 40(1), 34–69. <https://doi.org/10.1080/02702711.2018.1555361>
- Solomon, C., Lutkus, A. D., Kaplan, B., & Skolnik, I. (2004). *Writing in the nation's classrooms. Teacher interviews and student work collected from participants in the NAEP 1998 Writing Assessment*. ETS-NAEP Technical and Research Report 04-R02. ETS. <https://www.ets.org/Media/Research/pdf/ETS-NAEP-04-R02.pdf>
- Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4), 66–78.
- 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. <https://doi.org/10.1177/0265532212442647>
- Uzun, N. B., Alici, D., & Aktas, M. (2019). Reliability of the analytic rubric and checklist for the assessment of story writing skills: G and decision study in generalizability theory. *European Journal of Educational Research*, 8(4), 169–180. <https://doi.org/10.12973/eu-jer.8.1.169>
- 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 practices* (Vol. 27, pp. 23–32). Brill.
- Weigle, C. S. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11(2), 197–223. <https://doi.org/10.1177/026553229401100206>
- Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.
- Wesdorp, H. (1981). *Evaluatietechnieken voor het moedertaalonderwijs* [Evaluation techniques for the mother tongue education]. Stichting voor Onderzoek van het Onderwijs.