Psychological Test and Assessment Modeling, Volume 64, 2022 (3), 272-303

Meeting the Challenge of Assessing (Students´) Text Quality: Are There any Experts Teachers Can Learn from or Do We Face a More Fundamental Problem?

Ann-Kathrin Hennes[1], Barbara Maria Schmidt[2], Takuya Yanagida[3], Igor Osipov[4], Christian Rietz[5], Alfred Schabmann[6]

Abstract

Despite the importance of writing texts in school, teachers´ competence in assessing the quality of students' texts seems to be limited with respect to interrater reliability, i.e. objectivity. However, it is unclear whether the reason for this lies in the challenging task itself (assessing text quality) or is a matter of teachers' lack of expertise (which could be improved by better teacher training). In this study, groups of presumed experts, teachers, and novices rated the overall quality of 20 students´ texts. In addition, they rated the importance of different component properties of texts for text quality assessments. Their ratings of text quality/importance of criteria were compared within the framework of the expert-novice paradigm. A many-facet Rasch model analysis indicated that neither teachers nor any of the other expert groups met predefined expertise criteria. All groups' diagnostic competences were comparable to novices' competences. We argue that more effort must be undertaken to identify manifest criteria that define good texts and are suitable for use in school.

---

[1] Department of Therapeutic Education, Faculty of Human Sciences, University of Cologne, Germany. Ann-Kathrin Hennes: https://orcid.org/0000-0001-5526-9574. *Correspondence concerning this article should be addressed to* Ann-Kathrin Hennes, Department of Therapeutic Education, University of Cologne, Klosterstraße 79b, 50931 Cologne, Germany, Email: ann-kathrin.hennes@uni-koeln.de

[2] Department of Therapeutic Education, Faculty of Human Sciences, University of Cologne, Germany. Barbara Maria Schmidt: https://orcid.org/0000-0002-9167-0442

[3] Department of Applied Psychology: Work, Education, and Economy, Faculty of Psychology, University of Vienna, Austria. Takuya Yanagida: https://orcid.org/0000-0001-9052-4841

[4] School of Education, University of Wuppertal, Germany. Igor Osipov: https://orcid.org/0000-0002-1673-6894

[5] Department of Educational Science, Faculty of Educational and Social Sciences, University of Education Heidelberg, Germany. Christian Rietz: https://orcid.org/0000-0002-7057-4937

[6] Department of Therapeutic Education, Faculty of Human Sciences, University of Cologne, Germany. Alfred Schabmann: https://orcid.org/0000-0001-8523-9747

Keywords: assessing text quality; teachers' competences; many-facet Rasch model analysis; composition competence; expert-novice paradigm

Competence in composing a "functional text" is an important predictor of school success (Crossley & McNamara, 2016; Feenstra, 2014; Graham & Perin, 2007; Koster et al., 2015; National Commission on Writing, 2003; Becker-Mrotzek, 2014). Students often have to demonstrate their learning outcomes in written exams (MacArthur et al., 2016). Additionally, good writing helps students structure new information (Graham & Hebert, 2011; Linnemann & Stephany, 2014), which in turn is an important prerequisite for learning (Farnan & Fearn, 2008; Pohl & Steinhoff, 2010; Shanahan, 2006). Later, in the world of work, almost all jobs demand a minimum level of composing competence (CC) (Bach et al., 2016; McNamara et al., 2019).

Given its importance, imparting CC within writing instruction is a necessary part of any school curriculum (National Commission on Writing, 2003; Hodges et al., 2019). However, instruction should be based on students´ individual learning needs (Graham & Perin, 2007; Graham et al., 2012; Beck et al., 2018), which implies that teachers must have differentiated knowledge of students' individual strengths and weaknesses in CC and be able to assess students' CC accurately.

In the school context, students´ composition competence is typically assessed by measuring text quality (Feenstra, 2014; Philipp, 2015). To do so, teachers usually start by making a holistic judgment (in the form of an "overall impression"; Eckes, 2008, 160) and then (potentially) apply various more analytical criteria (Skar & Jølle, 2017). However, research has shown that teachers´ assessments of students´ texts might have poor interrater reliability (e.g. Birkel & Birkel, 2002; Cooksey et al., 2007; Skar and Jølle, 2017; Wyatt-Smith et al., 2010) and thus lack objectivity and validity.

Amongst others, research has identified the following factors that might impair the reliable (and in turn valid) assessment of students´ texts by teachers.

(1) Teachers might interpret and/or weight criteria inconsistently (Eckes, 2008), which means that they do not always measure the same text properties in the same manner when assessing text quality (Cooksey et al., 2007; Leckie & Baird, 2011; Murphy & Yancey, 2008; Olinghouse et al., 2012).
(2) Teachers tend to primarily use criteria that are easy to apply and have some kind of "face validity". Research has shown that teachers frequently examine very basic attributes of a text (e.g. orthography, grammar and vocabulary, which are not sufficient for a good text), rather than higher-order criteria that are more closely related to text quality and composing competence, such as text coherence (Birkel & Birkel, 2002; Rezaei & Lovorn, 2010; Vögelin et al., 2018; Vögelin et al., 2019; Smith et al., 2006). As a consequence, assessments say very little about students' composition competence and differ markedly from each other.

(3) Teachers frequently use their classes' overall performance level as an anchor for their ratings (Cumming et al., 2001; Cumming et al., 2002; Ingenkamp & Lissman, 2008; Madelaine & Wheldall, 2005) and compare students' texts with each other (Skar & Jølle, 2017). This works fairly well for ranking texts within classes (Cooksey et al., 2007), but is highly insufficient for assessing the "absolute" quality of a text (i.e. the quality of a text in relation to all texts in a given population of students). Depending on the class-average level, students with the same ability level can be characterized as "poor" or "good" writers. In particular, the performance level of a given class it is not sufficient to accurately identify struggling students, which compromises intervention (Boone et al., 2018; Marsh, 1987; Trautwein et al., 2006.

The reported findings suggest that teachers' assessments of text quality appear to be of poor reliability, i.e. teachers seem to lack expertise in diagnosing composition competence. However, to our best knowledge, there are no data comparing teachers' diagnostic competence to those of other potential "experts". Therefore, it is unclear whether teacher simply lack expertise (which would be the case if professions exist who already have good assessment competences to learn from) or if all groups of potential experts face a more fundamental problem with assessing text quality. The latter is quite conceivable, since criteria mentioned in the scientific literature (e.g., paying attention to the audience's needs; cf. Hennes et al., 2018; Knopp et al., 2014; MacArthur & Graham, 2016) are far from being well-operationalized (Knoch, 2011; Lumley, 2002; Todd et al., 2004).

Searching for Experts – the expert-novice paradigm

A promising approach to address the issue of objective, reliable and valid text assessments is the expert-novice paradigm (Voss, Fincher-Kiefer, Green & Post, 1986). This paradigm originally comes from creativity research (Amabile, 1982), but has also been applied in research on teachers' competencies (e.g. Bromme, 2008). By applying the expert-novice paradigm as part of the consensual assessment technique (Amabile, 1982), creativity research provides a framework for assessing complex constructs when no strong (i.e. manifest and objective) criteria are available – as is the case with text quality. The basic assumption underlying this approach is that experts in assessing creative products – as a result of their experience and reputation in the field – have implicit knowledge of key factors and criteria for good products in their domain of expertise (domain knowledge; e.g. Kaufman et al., 2013). The expert-novice paradigm assumes that expertise results from a person's strong and generally accepted contribution to the field of interest, which she/he can rely on when evaluating creative products (e.g. Kaufman et al., 2008; Amabile, 1982).

According to the expert-novice paradigm and findings from different creative areas, experts differ from novices mainly in two important points:

Criticalness: First of all, novices tend to give the same products better evaluations (higher ratings) than experts (Kaufman et al., 2008; Kaufman et al., 2009). Consequently, novices' ratings are less critical than experts' ratings.

Consistency: Second, experts' ratings are consistent. Experts evaluating the same product will largely be in agreement with each other (Amabile, 1996; Baer et al., 2004). On the contrary, novices' ratings are quite inconsistent (Kaufman et al., 2008).

Furthermore, the expert-novice paradigm assumes that (implicit) knowledge of relevant assessment criteria makes experts able to give consistent judgements. As previous research shows, knowing and consistently applying relevant criteria is one possible explanation for the high degree of consistency in experts' ratings (Eckes, 2008; Jonsson & Svingby, 2007; Weigle, 1994).

Writing research so far has not applied the expert-novice paradigm as used in creativity research. To date, there is only one older study by Diederich et al. (1961) comparing the consensus in assessing college students' text quality (short argumentative papers about different topics) in different occupational groups (English teachers, social scientists, natural scientists, lawyers, writers & editors, and business executives). In this study, the authors found that the correlations between raters' assessments (as a measure of consistency) were low in all groups (median correlation = .31), with English teachers having the highest median correlation of .41 and business executives the lowest median correlation of .22. However, it must be noted first that the comparison groups in Diederich et al.'s (1961) study included only one group (editors & authors) besides teachers to which special expertise in composition competence or text quality assessment might be attributed, but for which no information is available on their concrete field of expertise. Second, the evaluated texts were written in response to different prompts, i.e. differences in contextual factors might at least partially explain the different foci in different individuals' assessments.

However, if one goes beyond writing research and looks at findings from creativity research, it becomes apparent that Kaufman and colleagues were indeed successful in identifying experts who gave critical and consistent text quality ratings (compared to novices) to poems and fiction stories written by college students (cf. Kaufman et al., 2008; Kaufman et al., 2009).

Building upon these results, we continue these efforts by systematically searching for people who might already have the expertise necessary to assess basic aspects of text quality objectively and reliably, which is a mandatory (though not a sufficient) prerequisite for validly assessing text quality (Earle, 2020; Eckes, 2017; Huot, 2002; Lienert & Raatz, 1998; Moosbrugger & Kelava, 2012).

Research Questions

This study on the one hand explicitly raises the question of **whether any experts in assessing (students') text quality exist at all.** To answer this fundamental research question, the extent to which (possible) experts exhibit characteristics of diagnostic expertise compared to novices will be determined.

For this purpose, the following questions are examined:

(1) Does at least one of the (potential) expert groups – compared to novices – give critical text quality ratings?
(2) Does at least one of the (potential) expert groups – compared to novices – give consistent text quality ratings?

On the other hand, this study aims to answer the question of **how serious teachers' lack of expertise really is.** Therefore, teachers' diagnostic competences will be compared to those of (possible) experts on the one hand and novices on the other hand in order to position them on a continuum between possible experts and novices.

Concrete research questions are:

(3) How critical are teachers' ratings of text quality compared to the ratings of (potential) experts and novices?
(4) How consistent are teachers' ratings of text quality compared to those of (potential) experts and novices?

Furthermore, this study aims to **search for the main text assessment criteria experts use** and to examine **how consistently different assessment criteria are assessed as relevant by experts** (in comparison to novices and teachers). This rather exploratory process can provide important first impulses for the establishment of a valid "yardstick" for measuring text quality (cf. Cumming et al., 2002; Kaufman & Baer, 2012).

Concrete research questions are:

(5) Which criteria do the different groups consider highly relevant for assessing text quality?
(6) Are there criteria that are particularly important for all (potential) experts?
(7) How consistent are assessments of the importance of various text assessment criteria within each group?

Methods

Sample

We compared N=175 teachers', experts' and novices' evaluations of students' compositions.

Teachers: The group of teachers comprised N = 51 in-service primary school teachers who taught Grade 4 German at the time of the study or had taught it within the last three years (this was a selection criterion for participating in the study because the stimulus texts were written by fourth graders). On average, the participating teachers had worked as a teacher for 12.4 years (*SD* = 10.0), their mean age was 38.9 years (*SD* = 11.0), and 94.1% of them were female.

Corresponding to a study by Kaufman et al. (2009) focusing on the evaluation of narratives (written by university students) that defined both professional writers who published creative texts (see also Diederich et al., 1961) and scientists in the field of (English) language as experts, we identified two groups of experts:

Experts 1: The first expert group comprised N=31 editors & authors who edit/write books for children and adolescents. Of these, 87.1% were female. The mean age was 42.3 years (*SD* = 10.5); 35.5% worked as editors, 29% worked as authors and 35.5% worked as both editors and authors. Their average work experience was 10.8 years (*SD* = 6.8), and 56.5% (*SD* = 34.7) of their working time on average was spent on children's and youth literature. The focus on children's and youth literature was added as a criterion to create a more homogenous group than in previous studies (Diederich et al., 1961) and better match the participants' field of work/expertise to the context of the present study.

Experts 2: The second expert group consisted of scientists in the field of writing development and writing didactics. The sample consisted of 24 persons (60.0% female) with a mean age of 39.6 years (*SD* = 9.0). In this sample, 32.0% were doctoral students, 44.0% were post-doctoral researchers, and 24.0% were professors. Their average work experience in research was 11.0 years (*SD* = 7.7), and they spent on average 36.8% (*SD* = 21.5) of their working time on texts written for or by children and adolescents. All participants' main research areas were related to writing assessment. This group was chosen as an expert group because we assume that based on their experience (e.g. in the construction of rating scales for assessing text quality), they have deep knowledge of the text quality construct, the writing process and the state of research in the field of writing assessment.

Novices: The novice raters consisted of 69 university students (91.3% female) studying psychology or rehabilitation sciences. Their mean age was 24.4 years (SD = 3.3). It was ensured that no participants in this group had any prior knowledge in the field of text assessment or any experience in rating (students') texts. Those who had already completed a university course on assessing (students') text quality or adjacent topics

and/or who had worked as authors/editors alongside their studies were removed from the sample. Moreover, we assume that university students are broadly comparable to the other groups in terms of socio-economic and cognitive variables, which is important in order to ensure that the results regarding possible group differences are not driven by variables that we know have an influence on raters' text quality assessments (e.g. reading abilities; Schoonen et al., 1997), but which are not considered part of expertise.

Procedure for Recruiting Experts and Novices

We informed 10 of the participating *editors/authors* about the study through a German association of self-employed editors and authors and invited them to participate (Association of Freelance Editors; https://www.vfll.de/). The remaining 21 participants were selected via an Internet search and contacted personally by e-mail. All editors/authors were paid an expense allowance that they themselves specified (amounts paid varied between 0 and 130 Euro).

Scientists were selected via an intensive internet search (university websites) and a review of the relevant national (German) literature on writing. Potential participants were contacted personally via e-mail.

In order to recruit *teachers* for the study, we informed primary school principals about the study via e-mail and phone and asked them to encourage teachers in their schools to participate.

Novice raters were recruited in courses they attended at university.

Raters in all groups were informed about the aim of the study. All raters participated voluntarily and gave their consent for their data to be used for research purposes.

Instruments

Participants had to assess 20 texts of different quality. The online survey tool *soscisurvey* (www.soscisurvey.de) was used to host the text assessment and related (demographic) questionnaires. Participants were invited to participate in the study via e-mail and received a link and password enabling them to access the survey. The study was conducted between August 2016 and June 2018.

The Text Corpus

The study design had all raters assess the quality of the same 20 stimulus texts. In order to guarantee variance with regard to text quality, a representative sample of 20 texts was drawn from a larger corpus of texts. The text corpus used for this sampling procedure was generated by asking 401 fourth graders in North Rhine-Westphalia (Germany) from nine different schools (and 21 different classes) to write a short narrative text based on a set of six sequential pictures (see Figure 1). To standardize the content of the evaluated texts (and its influence on text assessment), all writers were given the same task with the same six pictures in the same order.

The task itself was selected in accordance with the German primary school curriculum, which explicitly mentions writing a story based on a set of pictures (e.g. Ministry for School and Further Education of North Rhine-Westphalia, 2012).

Figure 1
Story in pictures "Papa Moll goes paragliding" (©2010, Orell Füssli Sicherheitsdruck AG, Globi Verlag, Imprint Orell Füssli Verlag, Zürich, Picture from Papa Moll, Volume 10, PM the sporting ace)

To prepare the texts for rating, all students' (handwritten) texts were typed into a word processor and were corrected for spelling errors. However, the other parts of the text (e.g. headings and paragraphs) were kept unchanged. Next, we drew a random sample of 201 texts (from the existing 401 texts). These 201 texts were then rated anonymously in the D-PAC software (Lesterhuis et al., 2017) using the comparative judgement (CJ) method (Pollitt, 2012), which is well-established in writing research (c.f. van Daal et al., 2016). A sample of N = 47 students at the University of Cologne (teacher education students with a bachelor's degree who were attending a course on reading and writing skills at the time of the survey; mean age 25.1 years; 87.2%

female) were asked to rate the 201 texts by comparing them two at a time and determining which text was "better". In this way, each of the 201 texts was compared an average of twenty times to another text. These data were then used to calculate a logit value for every text (Bradley & Terry; 1952); the logit values calculated exhibited a satisfactory internal consistency (scale separation reliability = .81). In a final step, we divided the texts into 20 percentile groups based on their logit values and randomly selected one text from each percentile group.

Text Rating Procedure

Participants were asked to rate the 20 texts globally on a 6-point scale from 1 (poor quality) to 6 (high quality).

Before rating the texts, participants were informed of the writers' age and grade, as well as the instructions of the writing task the students had been given and how spelling errors had been handled as described above. No specific information about the authors' identity (e.g. gender or national origin) was given.

Over the course of their ratings, they saw 20 questionnaire pages, which all had the same structure. At the top of the page was the question: "How would you assess the overall quality of the presented text?". Below, one of the 20 stimulus texts was presented, and the six-point rating scale was depicted at the bottom of the page. Participants made their ratings by clicking on the corresponding point on the rating scale. The stimulus texts were presented in random order. The raters were not able to directly compare the texts to each other (e.g. by browsing back and forth). In this way, the raters should have been forced to assess the "absolute" quality of each text.

Assessment of the Importance of Individual Criteria

After completing all 20 global text ratings, participants were asked to rate four dimensions of composition competence according to their importance for their own decisions about general text quality. The dimensions to be evaluated were:

(1) paying attention to the audience's needs (e.g. Hayes, 2012)
(2) information management (e.g. Bereiter & Scardamalia, 1987; Hayes, 2012)
(3) knowing and deploying an appropriate text pattern (e.g. Heinemann, 2000)
(4) establishing coherence (e.g. MacArthur et al., 2019)

Each dimension was captured with a set of criteria (see Appendix B). All criteria were chosen based on an intensive literature search for text quality criteria and a comparison of different analytical rating scales used in research (e.g. Eckes, 2008; Hachmeister, 2019; Nussbaumer, 1991; Thonke et al., 2008). Some criteria were revised based

on a pilot survey of teachers conducted prior to the study to determine which criteria they consider when assessing text quality. This survey was conducted to find alternatives to academic formulations and thus to generate criteria that are equally understandable for all groups. In addition, five criteria (theoretically) related to lower-order writing skills (e.g. spelling; cf. Sturm & Weder, 2016) were presented.

Participants assessed the importance of each criterion for their evaluations on a 6-point rating scale (1=not important at all; 6 =very important).

Statistical Analysis

Using the Facets software, we conducted a many-facet Rasch model analysis (MFRA; Linacre, 2019) to answer our research questions.

MFRA is based on the rating scale model for polytomous responses proposed by Andrich (1978). It allows for modeling psychometric features of a rating (e.g. the evaluation of a text) as well as different aspects of rater behavior, rater effects and rater characteristics simultaneously (Wu, 2017).

Two different models were calculated based on a joint maximum likelihood estimation (Eckes, 2020). The first model (Model A) analyzed the global text quality ratings, while the second model (Model B) analyzed the personal importance of the assessment criteria for each rater. The basic equation for both models is given by:

$$log\left(\frac{p_{v,j(g),g,k}}{p_{v,j(g),g\ (k-1)}}\right) = \vartheta v\ -\ \alpha j(g)\ -\gamma g\ -\ \tau k$$

Specification of Model A: In Model A, $p_{v,j,k}$ stands for the probability of obtaining score *k* (e.g. "3" on the 6-point rating scale) for a particular text (v) evaluated by a particular rater (j). Similarly, $p_{v,j,(k-1)}$ is the probability of obtaining score k-1 (e.g. score "2" in the above example). The ratio of the two probabilities refers to the chance (odds) of a rater (j) assigning a particular score k (e.g. "3") compared to his or her chance of assigning score k-1 (e.g. "2"). In this model, the value of the odds ratio depends on three parameters: the quality parameter (ϑ_v) for a given text, reflecting its quality; the severity parameter for a rater (α_j), reflecting the rater's criticalness; and the threshold parameter (τ_k), which represents the transition point at which the probability of a text being rated in each of two adjacent categories is 50% (conditional upon it being in one of these two categories at all; Eckes, 2009). These parameters are crucial with respect to the empirical ordering of rating scale categories and hence the overall functioning of the rating scale. In a rating scale model (Andrich, 1978), the threshold parameters are set to be equal across all raters, which means that the rating scale functions in the same way for all raters. The addition of the group membership parameter (γ_g) allows for estimating group-specific rater parameters. A very convenient feature of the rating scale model is that all model parameters are calibrated on the same *logit* scale, which makes them directly comparable.

In the second model (Model B), the same model was fit to the data. This time, however, instead of a student text (see model A), a particular criterion (v) from the list provided in Appendix B was evaluated by a particular rater (j). The respective criteria were evaluated independently of concrete texts, so that no text facet was to be specified in this model. Therefore, in Model B, the value of the odds ratio depends on: the quality parameter (ϑ_v) for a given criterion, reflecting the importance raters assign to the criterion (i.e. rating of its relevance for assessing text quality); the severity parameter for a rater (α_j), reflecting the rater's global tendency to attribute more or less relevance to criteria; and the threshold parameter (τ_k) as described in Model A. Here again, a group membership parameter (γ_g) was added in order to estimate group-specific parameters.

In order to identify criteria that are particularly important for (potential) experts (research question 6), we performed an exploratory two-way group by criteria interaction analysis. For this purpose, the formula above was extended by a bias parameter (ϕ_{vg}), which represents the interaction between the facet criteria and the facet rater group.

MFRA results contain a comprehensive set of indices that allow conclusions to be drawn about the properties of the variables in the model (called facets) and possible rater effects. Indices relevant to answering the research questions posed in this article are briefly described in Appendix A.

Results

Participants´ Global Text Quality Ratings (Model A)

Model fit. Overall, Model A fit the data well. Only 3.9% of the absolute standardized residuals were equal to or greater than 2, and 0.6% were equal to/greater than 3. The data can be well described by the model.

Table 1
Rating Scale Category Statistics

Category	Absolute frequency	Relative frequency	Average Measure	Expected Measure	Outfit	Thresh-old	SE
1	198	6	-2.23	-2.10	0.9		
2	566	16	-1.30	-1.28	1.0	-2.75	0.8
3	914	26	-0.36	-0.41	1.1	-1.33	0.5
4	879	25	0.53	0.48	1.0	0.70	0.5
5	617	18	1.43	1.48	1.1	1.32	0.5
6	326	9	2.57	2.60	1.1	2.68	0.7

Note. Outfit is a mean-square fit index. Thresholds are calculated based on the Andrich rating model**.**

Rating scale quality. As shown in Table 1, there are more than 10 responses for each category, which means that, following Linacre (1999, 2004), the minimum requirements for a well-estimated and stable scale are fulfilled. Additionally, the response frequency across categories is regularly (unimodally) distributed, which indicates a regular usage of the rating scale and its categories (Eckes, 2015). Combined average measures increase monotonically with values on the six-point rating scale, so that it is safe to assume that higher ratings correspond to higher text quality. Furthermore, the scale has a nearly perfect model fit (expected measures and average measures correspond to each other very well, outfit statistics are <2 (Eckes, 2015). Moreover, Table 1 shows the threshold estimates and their standard errors for each category. It can be seen that these also increase monotonically with the categories; however, the criterion posed by Linacre (2004) that thresholds should increase by at least 1.4 logits and less than 5.0 logits is not met for each transition between categories. However, since this was the only limitation related to the quality of the scale, we kept it in its given form.

Facet 'text quality'. Homogeneity statistics showed that the presented 20 texts significantly differed (χ^2 (19) =3728.3; p<.001) and can be clearly distinguished from each other (separation reliability = 1.00) in terms of quality. These results imply that the selection of 20 stimulus texts varying in quality was successful.

Participants' criticalness when assessing text quality (research questions 1 & 3). The severity of teachers', experts' and novices' ratings were compared to each other. Overall, raters' severity (here: criticalness) ranged from -2.10 (mildest rater) to +2.07 (most critical rater) on the logit scale (see variable map, Appendix C). Group differences in severity did not reach overall statistical significance (χ^2 (3) = 3.9, p=.27), i.e. no group is particularly critical. On a descriptive level, however, it can be stated that the group of scientists is the most critical, followed by editors & authors (see table 2).

In contrast to the lack of group differences, the fixed chi-square values were significant for every group, indicating considerable differences in criticalness within groups (Table 2). Separation reliability was (equally) high within all groups, and in every group, several distinct classes of raters (with different levels of criticalness) could be identified.

Table 2
Raters' Severity: Sample Size, M (SD) of Average Severity Logit Value, Homogeneity χ^2 (df), Separation Reliability, Class Separation (Number of Strata)

	Teachers	Scientists	Editors & Authors	Novices
Sample Size	51	24	31	69
Average Severity Logit Value	-.03 (.04)	.06 (.06)	.03 (.05)	-.05 (.03)
Fixed chi-square	325.8 (50)*	140.0 (23)*	273.9 (30)*	530.2 (68)*
Separation Reliability	0.85	0.83	0.89	0.88
Class Separation	2.38	2.25	2.90	2.69

Note. * p < .05

Consistency of Global Text Quality Ratings (research questions 2 & 4). In the MFRA framework, consistency can be assumed if an individual rater's assessments fit the underlying measurement model. First, raters' individual fit statistics within each group were categorized as fitting or not fitting the model, in accordance with Linacre & Wright (1994). Second, the percentage of raters within each group fitting (or not fitting) the model was calculated (Table 3).

Overall, the differences between groups in terms of raters' consistency did not reach significance for either infit [χ^2 (3) = 4.03; p=.26] or outfit statistics [χ^2 (3) = 3.70; p=.30], i.e. no group is particularly consistent.

Scientists exhibited the highest proportion of raters fitting the model. This was equally true for infit and outfit statistics. The percentage of scientists whose ratings did not fit

the model (i.e. whose ratings were inconsistent with those of the other group members) was just 12.5%, about half as high as for teachers. Teachers exhibited the second-highest proportion of raters fitting the model. However, more than 20% of teachers deviated from the model's expectations and thus gave ratings differing from those of the other teachers (i.e. were inconsistent). Editors & authors as well as novices had the lowest percentages of participants whose ratings fit the model.

Table 3

Percentage of Participants Whose Ratings Fit the Model (i.e. Who Exhibited a Satisfactory Model Fit of 0.4-1.2). Infit: Ratings of "Average" Texts; Outfit: Ratings of Extremely Poor or Good Texts

	Teachers	Scientists	Editors & Authors	Novices
Infit	76.0	87.5	68.0	66.5
Outfit	78.0	87.5	68.0	69.0

Note. See Appendix A for further explanation.

Relevance of Assessment Criteria for Global Text Quality Ratings (Model B)

Model fit. Model B fit the data well. Only 3.1% of the absolute standardized residuals were equal to or greater than 2, and 0.9% were equal to/greater than 3. Again, the data were well described by the model.

Table 4

Rating Scale Category Statistics

Category	Absolute frequency	Relative frequency	Average Measure	Expected Measure	Outfit	Threshold	SE
1	106	3	-0.29	-0.40	1.3		
2	274	7	-0.08	-0.12	1.1	-1.21	0.11
3	394	10	0.26	0.22	1.0	-0.32	0.6
4	759	20	0.53	0.61	0.9	0.24	0.5
5	1208	32	1.05	1.05	0.9	0.36	0.4
6	1080	28	1.59	1.55	1.0	1.41	0.4

Note. Outfit is a mean-square fit index. Thresholds are Rasch-Andrich thresholds. Thresholds are calculated based on the Andrich rating model.

Rating scale quality. As shown in Table 4, there are again more than 10 responses for each category, which means that, following Linacre (1999, 2004), the minimum requirements for a well-estimated and stable scale are fulfilled. However, the response frequency across categories is irregularly distributed (right skewed). Nevertheless, combined average measures increase monotonically with values on the six-point rating scale, so that it is safe to assume that higher ratings correspond to higher text quality. Furthermore, expected measures and average measures correspond to each other very well, and outfit statistics for each category are close to 1 (and for all categories <2; Eckes, 2015), which indicates that the scale has a good model fit. However, the criterion posed by Linacre (2004) that thresholds should increase by at least 1.4 logits and less than 5.0 logits is not met for any transition between categories (see Table 4). However, since the limitations of the scale are due to the fact that we mainly specified theoretically relevant criteria and did not assume a regular distribution of the criteria in terms of their relevance, we kept the scale in its given form.

*Facet 'importance of criteria'***.** Homogeneity statistics showed that the 25 presented criteria significantly differed in importance (χ^2 (24) =1618.2; p<.001) and can be clearly distinguished from each other (separation reliability = .98).

Importance participants assign to the given set of assessment criteria. Overall, raters' severity (here, tendency to assign more/less importance to the given set of criteria) ranged from -1.7 (considering the criteria to be more important) to +1.1 (considering the criteria to be less important) on the logit scale (see variable map, Appendix D).

The differences between groups in severity (research question 5) reached a significant level (χ^2 (3) =21.6, p<.001). Teachers as well as editors & authors (both with negative average severity logit values) consider the given set of text assessment criteria to be more important when assessing text quality. On the contrary, scientists' and novices' average severity logit values were positive, indicating that they assigned less importance **overall** to the given set of criteria.

The fixed chi-square values turned out to be significant for every group, indicating considerable differences in the relevance attributed to the criteria within groups. Separation reliability was high for all groups, indicating that raters can be distinguished from one another with respect to the importance they assign to the criteria. Additionally, more than one class (strata) of raters was identified in each group. However, scientists´ ratings of the importance of the given criteria were more similar to each other than those of the other rater groups.

Table 5
Importance of Criteria: Sample Size, M (SD) of Average Severity Logit Value, Homogeneity χ^2 (df), Separation Reliability, Class Separation (Number of Strata)

	Teachers	Scientists	Editors & Authors	Novices
Sample Size	45	20	30	58
Average Severity Logit Value	-.09 (.03)	.11 (.04)	-.07 (.04)	.05 (.03)
Fixed chi-square	180.8 (44)*	61.6 (19)*	179.9 (29)*	297.6 (57)*
Separation Reliability	0.80	0.69	0.84	0.83
Class Separation	1.99	1.48	2.32	2.17

Notes. For computational reasons, numerically smaller units with regard to severity logit values denote higher relevance; * $p < .05$

Particularly important assessment criteria (research question 5 & 6). Basically, the results of the two-way group by criteria interaction analysis indicated that about a third of the combinations of group and criteria were associated with substantial differences between observed and expected ratings. For each of the studied groups, significant interactions could be identified in terms of criteria that were rated as more important than expected, as indicated by the bias measure's positive sign.

Table 6
*Summary Statistics for the Exploratory Interaction Analysis: N (number of element combinations); large T-values (absolute t-values equal or greater than 2.00); t-values statistically significant at p < .05. * p < .05; M (means) and SD (standard deviations) of the respective t-values*

Statistics	**Criterion X Group**
N	100
% large t-values[a]	31
% sig. t-values[b]	31
Min-t (df)	-4,43 (28)
Max-t (df)	4,17 (44)
M	0,02
SD	1,86

For **teachers,** the identified criteria mainly refer to the text pattern dimension. They rated the following criteria as more important than expected: "Considering the writing assignment" (bias = 1.13 logits, SE = .27; t = 4.17), "Writing an introduction" (bias = 1.03 logits, SE = .27; t = 3.82), "Writing an ending (bias = 0.71 logits, SE = .28; t = 2.54)", "Rising suspense" (bias = 0.50 logits, SE = .18; t = 2.74) and "Complete presentation of events/information" (bias = 0.43 logits, SE = .19; t = 2.30).

Novices primarily rated criteria related to lower-order writing skills as more important than expected. Those are: "Spelling mistakes" (bias = 0.46 logits, SE = .11; t = 4.14), "Handwriting" (bias = 0.42 logits, SE = .11; t = 3.81), "No grammar mistakes" (bias = 0.27 logits, SE = .11; t = 2.40) and "Providing as much information as possible" (bias = 0.25 logits, SE = .12; t = 2.18).

For **scientists,** criteria which were rated as more important than expected are: "Making the text work as a whole" (bias = 1.12 logits, SE = .52; t = 2.17), "Considering the reader's perspective" (bias = 1.11 logits, SE = .24; t = 4.59) and "Using text structure to guide/support the reader" (bias = 1.05 logits, SE = .32; t = 3.26).

Editors & authors (second potential expert group) attribute more relevance than expected to the following five criteria: "Creating reasonable relationships between sentences" (bias = 1.12 logits, SE = .52; t = 2.17), "Writing an enjoyable text (considering the function of a narrative text)" (bias = 0.71 logits, SE = .23; t = 3.08), "Creating an emotional moment" (bias = 0.60 logits, SE = .24; t = 2.55), "Using direct speech" (bias = 0.60 logits, SE = .28; t = 2.16) and "Choosing words consistently" (bias = 0.55 logits, SE = .24; t = 2.27).

Consistency of importance ratings (research question 7). To determine how "consistently" the different groups assess the importance of the given set of text assessment criteria (research question 7), the consistency of teachers', experts' and novices' ratings were compared to one another. This analysis followed the same procedure as described above (see Model A).

For the outfit statistics, the differences in consistency between groups did not reach a significant level [χ^2 (3) = 5.97; p=.11]. However, significant differences in consistency between groups were found for the infit statistics [χ^2 (3) = 8.0; p=.046]. This means that groups do not differ in consistency when assessing the importance of particularly important or particularly unimportant criteria (compared to the raters' anchor). However, the groups' consistency differs when assessing criteria of "average" relevance (compared to the raters' anchor) whose classification is less clear. For editors & authors, the percentage of raters giving consistent ratings when assessing criteria of "average" relevance was considerably lower than in the other groups (with standardized residuals < 2.0). The percentage of teachers who rated the importance of the assessment criteria consistently (raters fitting the model) was highest, followed by the groups of scientists and novice raters. The latter two groups did not differ from each other.

Table 7
Percentage of Participants Whose Ratings Fit the Model (i.e. Who Exhibited a Satisfactory Model Fit of 0.4-1.2). Infit: Ratings of Criteria of "Average" Importance; Outfit: Ratings of Criteria That Are Particularly Important or Unimportant

	Teachers	Scientists	Editors & Authors	Novices
Infit	80.0	75.0	50.0	69.0
Outfit	82.0	75.0	57.0	72.0

Note. See appendix A for further explanation.

Discussion

Previous studies have shown that teachers do not have sufficient diagnostic competences in text evaluation, i.e. they lack the knowledge to appropriately design writing instruction. Based on these findings, the literature has often concluded that teachers should be better trained in applying important criteria and generating reliable (expert) ratings (Beck et al., 2018; Hodges et al., 2019; Penner-Williams et al., 2009). However, there is currently no generally accepted standard that could be used to train teachers in assessing text quality. Hence, our study goes somewhat beyond this line of argumentation by seeking to answer an important question, namely: Are there any experts in assessing text quality teachers can learn from, or do we face a more fundamental problem, i.e. a lack of experts who are able to assess text quality reliably?

To answer this question, this study applied the expert-novice paradigm. According to this paradigm, experts in assessing text quality (in contrast to novices) are characterized as (1) being critical in their assessments, and (2) being consistent in assessing texts' quality.

Comparing teachers to novices with regard to these expertise characteristics, our data confirm that teachers exhibit only weak characteristics of expertise beyond a novice level in assessing students´ written compositions. Firstly, teachers tend to be relatively mild in their ratings. There is no statistical difference between teachers' severity (criticalness) and that of novices. Secondly, teachers did not rate the texts with a fairly high level of consistency (according to mean-square fit statistics; Linacre, 2002); no statistically significant differences in consistency between teachers and novices could be found.

However, it must also be noted that the two groups of supposed experts do not differ from teachers or novices in their criticalness or consistency when assessing text quality based on our data either. Significant differences in raters' individual assessment

standards were found within each group. Consequently, not only teachers but also both supposed expert groups met none of the pre-defined characteristics of expertise.

Group differences were found in how important the different groups considered the text assessment criteria (aspects that should be related to text quality based on the literature). Scientist tended to rate criteria identified in theory/literature (e.g. Hennes et al., 2018) as critically relevant for text quality as more important than expected based on the other groups' ratings, i.e. they focus on the reader's perspective ("Using text structure to guide/support the reader", "Considering the reader's perspective") and text coherence ("making the text work as a whole"). Editors & authors also focus on text coherence ("Choosing words consistently", "Creating reasonable relationships between sentences") and additionally on criteria related to text function and text pattern elements ("Creating an emotional moment", "Writing an enjoyable text (considering the function of a narrative text", "Using direct speech"). This focus probably arises from the fact that they write or edit narratives for children and youth themselves. Teachers clearly focus on criteria that are part of writing instruction and are taught to be directly important for a prototypical story ("Writing an introduction", "Writing an ending", "Rising suspense", "Complete presentation of events/information"). Moreover, they rate the criterion "Considering the writing assignment" as more important than expected, which again shows that they have a clear focus on the requirements they set in writing instruction. Novices clearly rate criteria identified in theory/empirical literature as less relevant for text quality as more important than expected ("Spelling mistakes", "Handwriting", "No grammar mistakes", "Providing as much information as possible").

Taken together, the results regarding the importance of predefined criteria for text assessment fit the proposed expertise of the studied groups: scientists ascribe particularly high relevance to criteria critical for text quality. Nevertheless, based on our results, it seems that the experts do not agree with each other in a significantly better way than novices or teachers regarding the relevance of the criteria, nor do they give text ratings with a significantly higher degree of consistency or that are significantly more critical. However, it should be noted that the group of scientists was most critical on the descriptive level, and with approx. 88% of raters fitting the model, gave the most consistent text quality ratings by far.

In line with our findings, there is a discussion in the literature about whether it is even possible to reliably assess the quality of texts (Barkaoui, 2007; Van Gasse et al., 2019). For example, Van Gasse et al. (2019) state that "the idea that identical perspectives regarding the evaluation of text quality are a prerequisite for high-quality assessments is outdated" (p. 270). They argue that uniform evaluations are hardly possible due to the complexity and multidimensionality of the construct.

From our point of view, however, this argument must be critically questioned for two reasons: First, the argument that the complexity of the construct hinders its uniform evaluation might be valid with regard to a very high level of text quality, but not when it comes to assessing basic composition competencies that need to be taught in school. Furthermore, from a methodological perspective, it is crucial that evaluators provide

objective and reliable judgements (Lim, 2011; van Gasse et al., 2019; Woehr and Huffcutt, 1994), since a lack of both always compromises validity (Eckes, 2017; Filer & Pollard, 2000; Stobart, 2009), which in turn prevents the accurate identification of struggling writers. That is why more effort must be undertaken to identify minimal criteria that define functional texts and are suitable for use in school so that a criterion-based reference standard is available.

Limitations

Some limitations have to be considered when interpreting and discussing the data and results at hand.

First, the criteria examined in our study were predefined based on existing literature on the topic of text quality (assessment). However, it is possible that research has not yet identified all relevant criteria for assessing text quality. Since the present criteria were derived from the literature and thus oriented on already existing assessment criteria, they are not manifest and therefore might possibly have been interpreted differently by different persons. This could be a possible explanation for the high variance within each group. Hence, more studies should be conducted with the goal of defining composition competence as concretely and completely as possible.

Furthermore, the criteria used in this study can be criticized as largely school/teacher-related criteria because they were primarily formulated based on a pilot survey of teachers and existing rating scales from the instructional context. One could argue that different groups of experts use different criteria (Eckes, 2008; Sakyi, 2000) as a result of their different goals/the context in which they work (e.g. train children to write texts with a minimum level of suitability for daily use, evaluating the quality of texts for publication, evaluating the linguistic precision of a text). This critique suggests that a basic requirement for minimal (!) assessment criteria is that they be objective, i.e. they should function independently of the assessor and the assessment context - although they must of course be adapted to the specific writing task, writer's experience and genre. Criteria that meet these requirements do not yet exist, as stated at the beginning of the article, but are urgently needed.

Another limitation of the data is that the criteria ratings referred to the importance raters assigned to a set of predefined criteria in general, not with reference to a specific text. It is feasible that raters´ actual assessment practice deviates from their self-reports on the criteria they use. Further research should be conducted focusing on raters' actual application and weighting of assessment criteria when assessing the quality of concrete texts, i.e. how these criteria influence global text quality ratings. Such an analysis would also shed light on the extent to which the dimensions of composition competence described in the literature really do fully describe the construct. First studies from the ESL context (e.g. Barkaoui, 2010) represent a good starting point here and should be expanded upon.

A last limitation relates to the sample sizes. Although most groups met the sample size requirements imposed by JMLE estimation methods, some were quite small (www.winsteps.com/facetman/estimationconsiderations.htm). Therefore, it cannot be assumed that these particular samples are representative of the larger population of teachers, editors & authors, scientists and novices. Hence, it will be necessary to verify the present findings in future replication studies. In addition, it would certainly make sense to make the various groups even more homogeneous, e.g. in terms of their specific work focus or professional experience. In this respect, the data showed considerable differences between the members of some of the groups.

References

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5), 997-1013. https://doi.org/10.1037/0022-3514.43.5.997

Amabile, T. M. (1996). *Creativity in Context: Update to the Social Psychology of Creativity*. Westview.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573. https://doi.org/10.1007/BF02293814

Bach, R, Schmidt, B.M., Schabmann, A., & van Koll, S. (2016). Braucht mein Friseur wirklich Zirkel und Lineal? – Schulisches Basiswissen im Kontext der Ausbildungsreife. [Does my hairdresser really need compasses and ruler? - Basic school knowledge in the context of the training maturity.] *Heilpädagogische Forschung*, 42, 61-72.

Baer, R. A., Smith, G. T., & Allen, K. B. (2004). Assessment of Mindfulness by Self-Report. The Kentucky Inventory of Mindfulness Skills. *Assessment*, 11(3), 191-206. https://doi.org/10.1177/1073191104268029

Barkaoui, K. (2007). Participants, texts, and processes in ESL/EFL essay tests: A narrative review of the literature. *Canadian modern language review*, 64(1), 99-134. https://doi.org/10.3138/cmlr.64.1.099

Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing*, 27(4), 515-535. https://doi.org/10.1177/0265532210368717

Beck, S. W., Llosa, L., Black, K., & Anderson, A. T. (2018). From assessing to teaching writing: What teachers prioritize. *Assessing Writing*, 37, 68-77. https://doi.org/10.1016/j.asw.2018.03.003

Becker-Mrotzek, M. (2014). Schreibkompetenz. [Writing competence] In J. Grabowski (Eds.). *Sinn und Unsinn von Kompetenzen: Fähigkeitskonzepte im Bereich von Sprache, Medien und Kultur*, 51-72. Verlag Barbara Budrich. https://doi.org/10.2307/j.ctvddzg18.5

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written communication*. Routledge.

Birkel, P., & Birkel, C. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. [How do teachers agree on text assessment? A replication study of the investigation by Rudolf Weiss.] *Psychologie in Erziehung und Unterricht*, 49, 219-224.

Boone, S., Thys, S., Van Avermaet, P., & Van Houtte, M. (2018). Class composition as a frame of reference for teachers? The influence of class context on teacher recommendations. *British Educational Research Journal*, 44(2), 274-293. https://doi.org/10.1002/berj.3328

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39, 324–345. https://doi.org/10.1093/biomet/39.3-4.324

Bromme, R. (2008). Lehrerexpertise: Eine psychologische Konzeption für die Entwicklung und Erforschung des Wissens und Könnens von Lehrern. [Teacher expertise: A psychological concept for the development and exploration of teachers' knowledge and skills.] In W. Schneider & M. Hasselhorn (Eds.). *Handbuch der pädagogischen Psychologie*, 159-167. Hogrefe.

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation*, 13(5), 401-434. https://doi.org/10.1080/13803610701728311

Crossley, S. A., & McNamara, D. S. (2016). Say More and Be More Coherent: How Text Elaboration and Cohesion Can Increase Writing Quality. *Journal of Writing Research, 7*, 351-370. https://doi.org/10.17239/jowr-2016.07.03.02

Cumming, A., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework (TOEFL Monograph Series, MS-22).* Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86, 67–96. https://doi.org/10.1111/1540-4781.00137

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability*. ETS Research Bulletin Series, 1961(2), i-93.

Earle, S. G. (2020). Balancing the demands of validity and reliability in practice: Case study of a changing system of primary science summative assessment. *London Review of Education.*

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

Eckes, T. (2009). Many-facet Rasch measurement. In V. Aryadoust and M. Raquel (Eds.). *Quantitative Data Analysis for Language Assessment Volume I*, 53-176. Routledge. https://doi.org/10.4324/9781315187815-8

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement. Analyzing and Evaluating Rater-Mediated Assessments. 2nd Revised and Updated Edition*. Peter Lang.

Eckes, T. (2017). Rater effects: Advances in item response modeling of human ratings–Part I. *Psychological Test and Assessment Modeling*, 59(4), 443-452.

Eckes, T. (2020). Rater-Mediated Listening Assessment: A Facets Modeling Approach to the Analysis of Raters' Severity and Accuracy When Scoring Responses to Short-Answer Questions. *Psychological Test and Assessment Modeling*, 65(4), 449-471.

Eckes, T., & Jin, K. Y. (2021). Measuring rater centrality effects in writing assessment: A Bayesian facets modeling approach. *Psychological Test and Assessment Modeling*, 63(1), 65-94.

Farnan, N., & Fearn, L. (2008). Writing in the disciples: More than just writing across the curriculum. In D. Lapp, J. Foold & N. Franan (Eds.), *Content area reading and learning*, 403-424. Erlbaum.

Feenstra, H. (2014). *Assessing writing ability in primary education: on the evaluation of text quality and text complexity.* Universiteit Twente.

Filer, A. and Pollard, A. (2000) *The Social World of Pupil Assessment: Process and contexts of primary schooling*. Continuum.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99(3), 445-476. https://doi.org/10.1037/0022-0663.99.3.445

Graham, S., & Hebert, M. (2011). Writing to read: A meta-analysis of the impact of writing and writing instruction on reading. *Harvard Educational Review*, 81(4), 710-744. https://doi.org/10.17763/haer.81.4.t2k0m13756113566

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of educational psychology*, 104(4), 879-896. https://doi.org/10.1037/a0029185

Hachmeister, S. (2019). Messung von Textqualität in Ereignisberichten. In: I. Kaplan & I. Petersen (Eds.), *Schreibkompetenzen Messen, Beurteilen und Fördern*, 79-99. Waxmann.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, 29, 369-388. https://doi.org/10.1177/0741088312451260

Heinemann, W. (2000). Textsorte–Textmuster–Texttyp. Text-und Gesprächslinguistik. [Text genre - text pattern - text type. Text and conversational linguistics.] In K. Brinker, G. Antos, W. Heinemann & S. Sager (Eds.), *Text-und Gesprächslinguistik: ein internationales Handbuch zeitgenössischer Forschung*, 507-523. de Guyter.

Hennes, A-K., Schmidt, B. M., Zepnik, S., Linnemann, M., Jost, J., Becker-Mrotzek, M., Rietz, C., & Schabmann, A. (2018). Schreibkompetenz diagnostizieren: Ein standardisiertes Testverfahren für die Klassenstufen 4-9 in der Entwicklung. [Diagnosing writing skills: A standardised test procedure for grades 4-9 under development.] *Empirische Sonderpädagogik*, 3, 294-310.

Hodges, T. S., Wright, K. L., Wind, S. A., Matthews, S. D., Zimmer, W. K., & McTigue, E. (2019). Developing and examining validity evidence for the Writing Rubric to Inform Teacher Educators (WRITE). *Assessing Writing*, 40, 1-13. https://doi.org/10.1016/j.asw.2019.03.001

Huot, B. (2002). (Re)Articulating Writing Assessment for Teaching and Learning. *All USU Press Publications*, 137. https://digitalcommons.usu.edu/usupress_pubs/137

Ingenkamp, K., & Lissmann, U. (2008). *Lehrbuch der Pädagogischen Diagnostik* (Vol. 6). [Textbook of Educational Diagnostics.] Beltz Verlag.

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

Kaufman, J.C., Baer, J., Cole, C.J., & Sexton, J.D. (2008). A Comparison of Expert and Nonexpert Raters Using the Consensual Assessment Technique. *Creative Research Journal*, 20(2), 171-178. https://doi.org/10.1080/10400410802059929

Kaufman, J. C., Baer, J., & Cole, J. C. (2009). Expertise, domains, and the consensual assessment technique. *The Journal of creative behavior*, 43(4), 223-233. https://doi.org/10.1002/j.2162-6057.2009.tb01316.x

Kaufman, J.C., Baer, J., Cropley, D.H., Reiter-Palmon, R., & Sinnett, S. (2013). Furious Activity vs. Understanding: How Much Expertise Is Needed to Evaluate Creative Work? *Psychology of Aesthetics, Creativity, and the Arts*, *7*(4), 322-340. https://doi.org/10.1037/a0034809

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, 16(2), 81-96. https://doi.org/10.1016/j.asw.2011.02.003

Knopp, M., Jost, J., Linnemann, M., & Becker-Mrotzek, M. (2014). Textprozeduren als Indikatoren von Schreibkompetenz: Ein empirischer Zugriff. [Text procedures as indicators of writing competence: an empirical approach.] In T. Bachmann & H. Feilke (Eds.), *Werkzeuge des Schreibens: Beiträge zu einer Didaktik der Textprozeduren*, 111-128. Klett.

Koster, M., Tribushinina, E., De Jong, P. F., & Van den Bergh, H. (2015). Teaching children to write: A meta-analysis of writing intervention research. *Journal of Writing Research*, 7(2), 249-274. https://doi.org/10.17239/jowr-2015.07.02.2

Leckie, G., & Baird, J. A. (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. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Comparative judgement as a promising alternative to score competences. In: E. Cano & G. Ion (Eds.), *Innovative Practices for Higher Education Assessment and Measurement*, 119-138. IGI Global. https://doi.org/10.4018/978-1-5225-0531-0.ch007

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, 28, 543–560. https://doi.org/10.1177/0265532211406422

Linacre, J. M., & Wright, B. D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, 8(2), 350.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome*

Measurement, 3, 103–122.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2), 878.

Linacre, J. M. (2003). Size vs. significance: Standardized Chi-Square fit statistic. *Rasch Measurement Transactions*, 17(1), 918.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258–278). JAM Press.

Linacre, J. M. (2012). Many-facet Rasch measurement: Facets tutorial. *Retrieved March, 19*, 2019. https://www.winsteps.com/a/ftutorial1.pdf (last retrieved on, 19.03.2019)

Linacre, J. M. (2019). *A user's guide to FACETS: Rasch-model computer programs*. Winsteps.com

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse*. Beltz.

Linnemann, M., & Stephany, S. (2014). Supportive writing assignments for less skilled writers in mathematic classroom. In P. Klein, L. Boscolo, S. Kirkpatrick & C. Gelati (Eds.), *Studies in writing: Writing as a learning activity*, 6-93. Leiden: Koninklijke Brill NV. https://doi.org/10.1163/9789004265011_005

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

MacArthur, C. A., Graham, S. & Fitzgerald, J. (2016). *Handbook of writing research*. Guilford Press.

MacArthur, C. A., & Graham, S. (2016). Writing research from a cognitive perspective. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.). *Handbook of writing research*, 24-40. Guilford Press.

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing*, 32(6), 1553-1574. https://doi.org/10.1007/s11145-018-9853-6

Madelaine, A., & Wheldall, K. (2005). Identifying low-progress readers: Comparing teacher judgment with a curriculum-based measurement procedure. *International Journal of Disability, Development and Education*, 52(1), 33-42. https://doi.org/10.1080/10349120500071886

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79(3), 280–295. https://doi.org/10.1037/0022-0663.79.3.280

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, Justice & Language assessment*. Oxford: University Press.

Ministry for School and Further Education of North Rhine-Westphalia, 2012. *Richtlinien und Lehrpläne für die Grundschule in Nordrhein-Westfalen* [Guidelines and curricula for elementary school in North Rhine-Westphalia]. Ritterbach Verlag.

Moosbrugger, H., & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion*. Berlin, Springer.

Murphy, S., & Yancey, K. B. (2008). Construct and consequence: Validity in writing assessment. In C. Bazermann (Eds.), *Handbook of research on writing: History, society, school, individual, text*, 365-385. Taylor & Francis.

National Commission on Writing. (2003). *The neglected "r": The need for a writing revolution.* New York: College Entrance Examination Board.

Nussbaumer, M. (1991). *Was Texte sind und wie sie sein sollen. Ansätze zu einer sprachwissenschaftlichen Begründung eines Kriterienrasters zur Beurteilung von schriftlichen Schülertexten. [*What texts are and how they should be. Approaches to a linguistic justification of a criteria grid for the evaluation of texts written by pupils.] Tübingen: Niemeyer. https://doi.org/10.1515/9783111632308

Olinghouse, N. G., Santangelo, T., & Wilson, J. (2012). Examining the validity of single-occasion, single-genre, holistically scored writing assessments. In E. van Steendam, M. Tillema, G. Rijlaarsdam & H. van den Bergh (Eds.), *Measuring Writing, Recent Insights into Theory, Methodology and Practice*, 55-82. Brill. https://doi.org/10.1163/9789004248489_005

Orell Füssli Sicherheitsdruck AG (2010). Papa Moll, Band 10, *Papa Moll die Sportskanone* [Papa Moll - the sporting ace]. Globi Verlag.

Penner-Williams, J., Smith, T. E., & Gartin, B. C. (2009). Written language expression: Assessment instruments and teacher tools. *Assessment for Effective Intervention*, 34(3), 162-169. https://doi.org/10.1177/1534508408318805

Philipp, M. (2015). *Schreibkompetenz: Komponenten, Sozialisation und Förderung*. [Writing competences: components, socialisation and promotion.] A. Francke Verlag.

Pohl, T., & Steinhoff, T. (2010). *Textformen als Lernformen*. [Text forms as forms of learning.] Gilles & Francke.

Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2), 157-170. https://doi.org/10.1007/s10798-011-9189-x

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

Sakyi, A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Eds.), *Studies in Language Testing. Fairness and Validation in Language Assessment*, 129-152. Cambridge University Press.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing*, 14(2), 157-184. https://doi.org/10.1177/026553229701400203

Shanahan, T. (2006). Where does writing fit in Reading First? In C. Cummins (Eds.), *Understanding and implementing Reading First initiatives*, 106-115. International Reading Association.

Skar, G. B., & Jølle, L. J. (2017). Teachers as raters: Investigation of a long-term writing assessment program. *L1 Educational Studies in Language and Literature*, 17, 1-30. https://doi.org/10.17239/L1ESLL-2017.17.01.06

Smith, M., Cheville, J., & Hillocks, G. (2006). I guess I'd better watch my English. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing research*, 263-274. Guilford Press.

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research*, 51 (2), 161–79. https://doi.org/10.1080/00131880902891305.

Sturm, A., & Weder, M. (2016). *Schreibkompetenz, Schreibmotivation, Schreibförderung. Grundlagen und Modelle zum Schreiben als soziale Praxis*. [Writing competence, motivation to write, promotion of writing. Basics and models for writing as a social practice.] Kallmeier.

Thonke, F., Groß Ophoff, J., Hosenfeld, I., & Isaac, K. (2008). Kriteriengestützte Erfassung von Schreibleistungen im Projekt VERA. *Checkpoint Literacy*. Tagungsband, 2, 28-35.

Todd, R. W., Thienpermpool, P., & Keyuravong, S. (2004). Measuring the coherence of writing using topic-based analysis. *Assessing writing*, 9(2), 85-104. https://doi.org/10.1016/j.asw.2004.06.002

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006) Tracking, grading and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics, *Journal of Educational Psychology*, 98(4), 788–806. https://doi.org/10.1037/0022-0663.98.4.788

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2016). 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*, 1-16. https://doi.org/10.1080/0969594X.2016.1253542

Vögelin, C., Jansen, T., Keller, S. D., & Möller, J. (2018). The impact of vocabulary and spelling on judgments of ESL essays: an analysis of teacher comments. *The Language Learning Journal*, 1-17.

Vögelin, C., Jansen, T., Keller, S. D., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assessing Writing*, 39, 50-63. https://doi.org/10.1080/09571736.2018.1522662

Voss, J. F., Fincher-Kiefer, R. H., Greene, T. R., & Post, T. A. (1986). Individual differences in performance: The contrastive approach to knowledge. *Advances in the psychology of human intelligence*, 3, 297-334. https://doi.org/10.1016/j.asw.2018.12.003

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11, 197-223. https://doi.org/10.1177/026553229401100206

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of occupational and organizational psychology*, 67(3), 189-205.

Wright, M., & Masters, G. (2002). Justified criticism, misunderstanding, or important steps on the road to acceptance. In E. G. M. Weitekamp & H.-J. Kerner (Eds.), *Restorative justice: Theoretical foundations*, 50-70. Willan Publishing.

Wu, M. (2017). Some IRT-based analyses for interpreting rater effects. *Psychological Test and Assessment Modeling*, 59(4), 453-470.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy & Practice*, 17(1), 59-75. https://doi.org/10.1080/09695940903565610

Appendix A.

Explanations of MFRA indices

Severity logit values provide information about raters' assessment standards (e.g. their level of criticalness). Positive severity logit values indicate a rater's tendency to award lower ratings to a product/object (i.e. be critical), while negative logit values indicate the tendency to award higher ratings (i.e. be less critical or mild).

Fixed chi-square values are used to test whether the elements of a facet are from a single homogenous population (i.e. are statistically equivalent). Chi-square statistics are used to determine if parameters (e.g. raters' severity) can be fixed across all elements of a facet (Linacre, 2003; Linacre, 2012). These statistics therefore provide information on whether the members of a group share the same assessment standards (e.g. level of severity) or differ in these (in which case the chi-square test becomes significant).

Separation reliability coefficients provide information about the degree to which individual elements of a facet can be distinguished from each other (i.e. to what degree they differ). The higher the values, the clearer the differences (e.g. in raters' level of severity) between the individual elements of a facet.

Class separation coefficients specify the number of distinct classes ("number of strata"; Wright & Masters, 2002) into which the elements of a single facet can be divided. Elements within classes are similar, but there are marked differences between classes. Consequently, a one-class solution indicates that the differences between the individual elements of a facet are negligible (e.g. raters share the same assessment standards); the occurrence of more classes is indicative of nonhomogeneous facets.

Mean-square fit statistics provide information about how well individual elements of a facet fit the model's expectations. Within the MFRA, these fit statistics are used to analyze raters' judgment consistency (Linacre, 2002). Raters exhibiting a satisfactory model fit are assumed to give consistent ratings (i.e. they are in agreement with the model and therefore also with each other). Two different model fit indices can be calculated (comparable to item analyses based on the Rasch model): infit and outfit statistics. Both statistics refer to raters' behavior when assessing a product (e.g. a text). Infit statistics describe raters' behavior (i.e. consistent vs. inconsistent) when assessing products whose quality (specified in logits) is close to their severity parameter (maximum difference of 0.5 logits), i.e. the rater's anchor value of "average" quality. Conversely, outfit statistics describe raters' behavior when assessing very poor or very good products compared to their individual anchor value. Values for both statistics range from 0 to ∞ . Fit values of 1 represent a perfect model fit, while values greater than 1 indicate more variation than expected (e.g. a rater with a fit of 1.25 shows 25 % more variance in his/her ratings than expected), and values smaller than 1 indicate less variation than expected (e.g. a rater with a fit of 0.70 shows 30 % less variance in his/her ratings than expected). Depending on the direction of deviation from the

model, raters can be classified as ether “overfitting” the model (showing less variance in their ratings than expected and hence being too predictable) or “underfitting” the model (showing more variance in their ratings than expected and hence being too unpredictable; Eckes & Jin, 2021). In this study, the range of plausible (rater) fit indices was defined as 0.4-1.2, which according to Linacre & Wright (1994) is the spectrum of plausible fit indices for assessments in which consistency between raters is desired (as is the case for the given research question).

The size of a bias parameter provides information about the difference between the observed and the expected logit values. Dividing the estimate of the bias parameter by its standard error, a t-value is determined, which is tested for significance (here the hypothesis that there is no other source of error besides the general measurement error is tested; Eckes, 2015). T-values equal or greater than 2 indicate that the interaction under consideration contributes to the explanation of the model error and is thus relevant.

Appendix B.

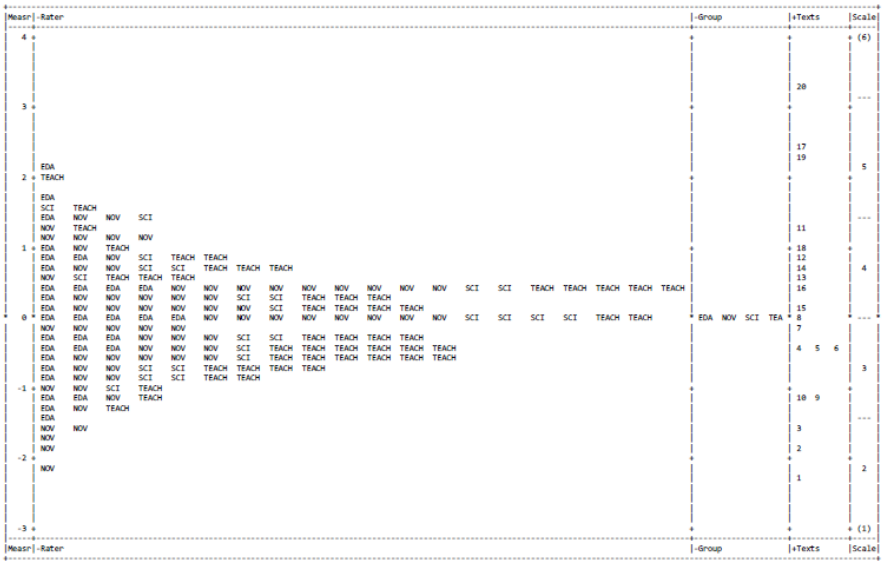
Dimensions of CC and corresponding criteria rated according to their relevance; criteria were numbered for MFRA (see parentheses).

Components of CC	Presented criteria
Information management	● Providing enough information to make plot comprehensible (1) ● Providing as much information as possible (2) ● Not providing too much information (3)
Deploying appropriate text pattern (narrative text)	● Writing an introduction (4) ● Writing an ending (5) ● Using tenses accurately (6) ● Using direct speech (7) ● Raising suspense (8) ● Creating an emotional moment (9) ● Using text structure to guide/support readers (10)
Paying attention to the audience's needs	● Considering readers' perspective (11) ● Presentation of events/information in a transparent order (12) ● Complete presentation of events/information (13)
Establishing coherence	● Using structure elements (14) ● Choosing words consistently (15) ● Making the text work as a whole (16) ● Creating reasonable relationships between sentences (17) ● Creating reasonable relationships between words (18) ● Considering the writing assignment (19) ● Writing an enjoyable text (considering the function of a narrative text) (20)
Others	● No spelling mistakes (21) ● No grammar mistakes (22) ● Text length (23) ● Vocabulary (24) ● Handwriting (25)

Note: Other operationalization possibilities would certainly also be conceivable; the formulations used here were (mostly) generated from qualitative surveys in which teachers themselves named criteria they use in assessing text quality. The presented criteria were assigned to the components of CC on the basis of theory. Due to theoretical overlap among the components of CC, the categorization may be occasionally unclear/ambiguous.

Appendix C.

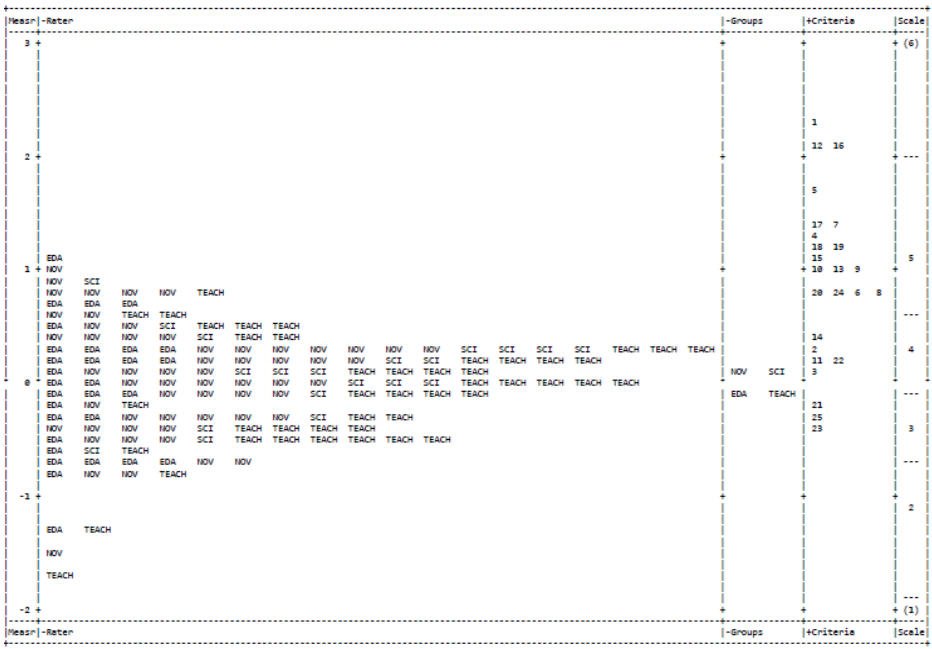
Variable map for MFRA Model A



Note: The first column shows the logit scale. Values on this scale indicate raters' severity and the global quality of the assessed text products. The second column shows the estimated distribution of raters' severity. Each abbreviation represents 1 rater (TEACH = teacher; SCI = Scientist; EDA = editors & authors; NOV = novice). In the third column, rater groups are added but not calibrated, since group membership was used only as a dummy facet in the analysis. The fourth column shows the estimated text quality parameters. Each number represents one text (1 to 20). The fifth column shows the 6-point rating scales on the logit scale so that text quality can be defined according to the 6-point rating scale. Threshold measures for the rating scale are represented by horizontal dashed lines.

Appendix D.

Variable map for MFRA Model B



Note: The first column shows the logit scale. Values on this scale indicate raters' severity and the relevance raters assigned to the given text assessment criteria. The second column shows the estimated distribution of raters' severity. Each abbreviation represents 1 rater (TEACH = teacher; SCI = Scientist; EDA = editors & authors; NOV = novice). In the third column rater groups are added but not calibrated, since group membership was used only as a dummy facet in the analysis. The fourth column shows the estimated relevance of the assessment criteria. Each number represents one criterion (1 to 25; see Appendix A). The fifth column shows the 6-point rating scales on the logit scale so that the criteria's relevance can be defined according to the 6-point rating scale. Threshold measures for the rating scale are represented by horizontal dashed lines.