Article

Written Communication
2024, Vol. 41(2) 203–229

DOI: 10.1177/07410883231222952
journals.sagepub.com/home/wcx


A Direct Functional Measure of Text Quality: Did the Reader Understand?

Joachim Grabowski[1] and Moti Mathiebe[1]

Abstract

Assessing text quality as an indication of underlying skills still remains challenging; irrespective of the approach, many studies struggle with reliability or validity problems. If writing is considered problem-solving, a report must make the reader understand the described situation and call for its mental reconstruction. Therefore, text quality may not only comprise linguistic aspects but also the cognitive-functional power of a text. The presented study aims at exploring the functionality of students' reporting texts in relation to general text-quality measures, using a corpus of accident reports written by German fifth- and ninth-graders (n = 277) prompted by a pictorial stimulus of a bike accident scenario. An online tool was developed in which 277 university students graphically reenacted the situation from one respective text according to the existence, position, and color of the involved elements. Thereafter, the match of the resulting spatial reconstructions with the original situation was assessed by two raters. While most subscales showed sufficiently high interrater reliabilities, the aggregated functionality score ($\alpha = .74$) had medium-high correlations with other text-quality ratings and was comparably dependent on grade, education level, and linguistic family background. However, the correlational pattern, regression analysis, and factor analysis showed that the functionality score also contributed unique portions of variance to the assessment of writing skill that were not

[1]Leibniz University of Hannover, Germany

Corresponding Author:
Joachim Grabowski, Institute of Psychology, Leibniz University of Hannover, Im Moore 11, Hannover 30167, Germany.
Email: grabowski@psychologie.uni-hannover.de

represented by rating measures. Moreover, the direct indication of whether a text allows for the reader's adequate cognitive representation is evident. Altogether, the approach of indicating text functionality through practical understanding offers a sound, though empirically laborious, alternative for text-quality measurement. Results are discussed with regard to the didactical strategy according to which students can improve their writing when they observe whether others can make use of their texts.

Keywords
reports, writing assessment, writing skill, web-based experiment, rating

Introduction

Assessing text-quality is one of the major challenging issues within writing research. However, the question of what constitutes a "good text" cannot be answered easily, and answers vary across disciplines and research fields. While researchers from cognitive psychology or psycholinguistics are interested in the writing processes that lead to textual products, developmental psychologists or writing didacticians focus on the question of how children acquire the underlying abilities to write a good text (see Neumann, 2012). Thus, different theories and perspectives on writing lead to different decisions concerning research design, measurement, data evaluation, and inferences in the assessment of writing. In typical educational studies, the obtained quality measures (e.g., scores or text length) are subsequently either related to other product or process characteristics (as in the case of correlational studies; e.g., Graham et al., 1997) or analyzed depending on systematically varied conditions or instructional measures (as in the case of experimental or intervention studies; e.g., Grabowski et al., 2010). Aside from research contexts, teachers also assess student texts in order to inform students about their individual achievements, and large-scale assessments (e.g., National Assessment Governing Board, 2010) describe the efficiency of the respective educational system. In any case, text-quality is operationalized by defined measures. The obtained measures are then interpreted within a theory framework with respect to the theoretical construct(s) they are assumed to indicate (cf. Van Steendam et al., 2012, for a comprehensive overview of writing measurement). This step of interpretation may pose further theoretical and methodological challenges with respect to the validity of such interpretations (Kane, 2013).

In this article, we describe a reference study and report on a follow-up study in which our main focus was on assessing the text-quality of individual

learners in a German secondary school classroom environment. In the reference study, we designed a study on the writing abilities of students at the beginning and at the end of German secondary school (i.e., Grades 5 and 9) attending three different school types (of lower, medium, and higher educational level; we will explain this differentiation within the German school system in the Methods section). About half of the students spoke only German at home; the other half had family languages different from German (i.e., linguistic migration biography [LMB]). These students had written a reporting text based on a pictorial stimulus.

Several measures on cognitive and linguistic abilities were taken in order to estimate the predictability of writing skill from these variables. The students' writing skills were indicated by quality measures of their texts. Since experts recommend using multiple variables for text-quality assessment (Schoonen, 2012), because of potential reliability (Diederich et al., 1961) and validity problems (e.g., Chan & Yamashita, 2022), multiple approaches were taken and combined in order to obtain a robust aggregated quality measure: one statistical parameter (text length) and different (either more holistic or more analytical) rating approaches. Indeed, the resulting text-quality scores turned out to be highly predictable from the collected prerequisite variables.

Given this reference study and its results (Grabowski et al., 2014, 2018), we designed a subsequent experimental study in which the reporting texts written by German secondary school students were subjected to a further, functionally very forceful, quality test: in the current study, we had university students—one student reader for each text—reconstruct the original situation described in a report. We then scored the achieved degree of similarity between the reconstruction and the original. We consider this method a direct measure of the immediate functional quality of texts because it requires the reader to construct an appropriate mental representation of the reported event and the related situation. After an analysis of the resulting scores according to grade, school type, and LMB of the original writers, we compared the obtained results to typical text-based quality ratings and analyzed how the different text-quality approaches may relate to one another and to the indication of the underlying skill to write a proper report.

Assessment of Text Quality: Text Properties, Judgments, and Text Functionality

Given the many different approaches to text-quality assessment, which will here be augmented by a further empirical facet, the question arises whether there is one unitary text quality or whether the notion and meaning of text

quality may vary according to the respective foci of the applied assessment methods. Answers to this question can be sought theoretically (e.g., Shermis, 2022) as well as empirically (e.g., Chan & Yamashita, 2022). We will provide a brief survey of the most typical approaches, while the specific procedures used for our analyses will be detailed in the Methods section.

Generally, there are two groups of approaches that can be used for indications of text quality: (a) determining (descriptive) features of the text or (b) obtaining (evaluative) judgments on text aspects. Both groups of measures can, in principle, be performed by human readers or researchers without technical tools (through operations like identifying, classifying, counting, or scoring). Meanwhile, there are also digital and automatic tools available for the most approaches and methods. It is debatable whether, and to what degree, the so-called automatic scoring of essays allows for meaningful judgments of text-quality aspects and whether it can, or should, replace human graders (overviews include Ke & Ng, 2019; Ramesh & Sanampudi, 2022; Shermis, 2022; most recently Vo et al., 2023).

(a) A wide range of studies have shown that there are certain statistical measures derived from texts that correlate with their overall quality scores. One of these measures is text length, that is, the overall number of words in a text (Pohlmann-Rother et al., 2016; Quasthoff & Domenech, 2016); another measure is lexical diversity, indicated, for example, by the *corrected type token ratio* (CTTR: Mathiebe, 2018; Olinghouse & Leaird, 2009) or the *measure of textual lexical diversity* (MTLD: Olinghouse & Wilson, 2013). Other studies examine the relation between text quality and certain means of academic language, like word frequency (Deno et al., 1982; Grobe, 1981; Mathiebe, 2018; Olinghouse & Leaird, 2009) or the proportion of content words within a text (Mathiebe, 2018; Olinghouse & Wilson, 2013). The CTAP-Tool (Common Text Analysis Platform; Chen & Meurers, 2016) offers an automatic analysis of many features that relate to the linguistic complexity of a text.

Even though such measures are generated in a reliable way, their validity remains disputable if they are considered only as an aggregate based on a single number or statistic. It seems unclear what these measures indicate beyond the mere description of the respective features of the linguistic surface of texts. An increasing use of the above-mentioned features (e.g., word frequency or proportion of content words) does not automatically lead to better texts—although, when writing skills are expanded and differentiated in school contexts, a higher amount of the respective linguistic text features is

often in line with a perceived better text. Nevertheless, these kinds of text-quality indicators do not allow for conclusions about their appropriate and functional application within the text and are not sufficient to draw any general inferences. This limitation shows that it is necessary to broaden the view on validity to a more complex concept. Kane (2013), for example, considers validity as a multifaceted concept that is context-dependent and asks for careful evaluation in terms of the plausibility of claims based on supposed indicators of text-quality.

Particularly, there appear to be at least two potential steps in the interpretation of text-quality measures. First, text quality may be seen as a linguistic construct, indicating characteristics that are attributed to the text itself, where a "good" text is described on aspects like content, structure, style, and formal correctness. Second, the quality of one or more written texts can be taken as an indication of the writing ability of the respective writer. Here, the measure indicates a psychological construct attributed, for example, to a learning student. Both kinds of assumed indications can be meaningful, under different empirical and theoretical conditions. With respect to the current debate on the use of large language models in educational contexts (Kasneci et al., 2023), a text created by AI (e.g., by chatGPT) can display high text-quality without any reference to the writing abilities of the person who requested the text from the software.

(b) The second common way of measuring text-quality is using judgments from readers. Texts may be considered communicative activities between writers and readers who are separated in time and space (Ehlich, 1984). Therefore, it is reasonable to involve the readers in the assessment of text-quality by asking for their evaluation. The applied rating methods vary between holistic and analytical approaches depending on the underlying theoretical assumptions (which include assumptions on the writing process as well as assumptions on the quality construct implemented in the respective task (cf. Shermis, 2022)); on the related academic disciplines and their traditions; but also on economic and temporal resources during the evaluation process (Schipolowski & Böhme, 2016). Moreover, the use and practicability of existing scoring schemes is usually restricted to certain genres, or even kinds of tasks. In contrast to the aforementioned descriptions of linguistic features, however, such ratings generally relate to aspects, or dimensions, for which it is more or less clear what constitutes better or worse characteristics of text quality. When holistic or overall ratings are obtained, for example, it is assumed that different aspects of a text that are related to its quality will not vary

(too) independently from each other. In didactical contexts, typical rating schemes (often called rubrics; cf. Imbler et al., 2023) refer to a componential score in which aspects like content, structure, style, and correctness are rated independently, and sometimes aggregated into a sum score, which then represents more than an overall impression. While such componential ratings are sometimes already called analytic, an analytic text-quality rating in its narrower sense refers to an elaborate linguistic model of text characteristics (refer to Nussbaumer & Sieber, 1994, for German texts) that allows to describe and evaluate a text on many specific aspects relevant for the respective genre (e.g., among other things, whether a letter contains a proper salutation formula, or an instruction uses the imperative). Basically, the further use and purpose of the obtained assessment plays a decisive role in the choice of the applied method. If a concluding assessment is expected, for example, in a leaving certificate or matriculation exam, a holistic score may be appropriate, while differentiated feedback on the diverse components of text quality may offer more help to learners who want to improve their writing skills in a classroom context.

All rating procedures have their strengths and weaknesses (Neumann, 2012), but they also are limited by the implied view of the reader compared to real and authentic readers (or better: *users* of the text). Reading a text for rating is different from reading a text for comprehension, or entertainment. In particular, raters who have to judge large amounts of texts gradually tend to focus only on the addressed rating categories, or simply follow the rating guidelines. They often automatically fragment the text according to the different rating aspects, which are by themselves not functionally based, for example, when the ways of expressing temporality within a report are determined. The text, as such and as a whole, is often lost out of sight, and it remains unclear whether the text is effective in overcoming the information gap between the reader and the writer. Thus, a further approach to the measurement of text quality is required that examines texts according to their effectiveness and that tests competent readers' overall comprehension as a function of the respective text. This view of text quality conforms to writing as a problem-solving process (Hayes & Flower, 1980).

In the writing as a problem-solving process view, the writer is supposed to solve a communicative problem by producing a text that fits the communicative needs of both the writer and the reader. Successfully composing a text that fits the communicative needs of the writer *and* reader is the reason why writing a text is a very challenging task. The writer must coordinate two representations of the text simultaneously, the text produced so far (What

have I written?) and the communicative target or intention (What do I want to say?). Additionally, the readers' perspectives and their prior knowledge must be considered (How does my reader interpret the text?). This third component is often the most difficult, because a (generalized) reader is hard to imagine. Particularly with writing assignments in school or university, writing instructions often lack descriptions of specified target readers. Moreover, many writers generally have problems anticipating the reader's needs and potential questions while reading their texts. Therefore, it seems appropriate to indicate the quality of a text by testing whether, or to which degree, the reader's needs are fulfilled and the text functions as a communicative activity. Even though in empirical writing research, where writing is most often considered as problem solving, the selected "problem solution" is rarely proven and put to the test. For didactical purposes, it can support the improvement of the learners' writing skills more than mere feedback on certain features of an already written text, if they become aware of how the readers of their texts fail, or succeed, with respect to the intended communicative goals (Rijlaarsdam et al., 2008).

A genre suitable to examine whether a text is written in a target-oriented communicative way is reports. In our study we use an accident report, where the course of events has to be explained by the writer such that the reader is able to understand the situation. Explaining the sequence that leads to the accident calls for restructuring the different representations in mind. The writer turns from a knowledge knower to a knowledge sender and has to verbalize this knowledge in a coherent way. At the same time, writers must follow the conventions of the respective text genre—in this case a report—which include the enclosed information as well as certain linguistic means. If the writer's description of the accident is successful, the reader should be able to reenact the situation on the basis of the composed text.

After we have described alternative approaches to obtain measures that represent the quality of a text, ending with a pragmatic criterion of immediate communicative success, we will subsequently concentrate on the empirical question of how to assess the last-mentioned functionality and effectiveness of a text. To that end, we will address the following research questions: (1) Does the quality of reporting texts, as measured through the readers' reconstructions of the described accident situation, systematically vary according to student writers' grade, school type, and family language? If the developed measure is valid and meaningful, it should, by and large, repeat the general result patterns known from other assessment approaches. (2) How does the functional measure correlate with other approaches to text-quality assessment? Does it simply capture the same aspects, or does it reveal some unique variance of text quality?

Table 1. Composition of Text Corpus According to Writer Characteristics.

	Educational Level of Secondary School Type			
Grade	Low (Hauptschule)	Medium (Realschule)	High (Gymnasium)	Total
Fifth grade	43 (17/26)	49 (12/37)	54 (29/25)	146 (58/88)
Ninth grade	39 (23/16)	40 (18/22)	52 (36/16)	131 (77/54)
Total	82 (40/42)	89 (30/59)	106 (65/41)	277 (135/142)

Note. Numbers within parentheses indicate students without (first number) and with (second number) linguistic migration biographies.

Methods

Reference Study

Participants. The reporting texts of this corpus were written by pupils from Grades 5 and 9. In Germany, Grades 5 to 10 typically constitute the lower secondary level (while Grades 11 to 12 or 13 form the upper secondary level). Germany has a tripartite secondary educational system where teaching is based on different curricula. Pupils within this study attended three different school types: Hauptschule (low level), Realschule (medium level), and Gymnasium (high level). In each school type, teaching is basically related to a specific target qualification (from apprenticeship to academic studies). At the end of lower secondary level, a general school-leaving certificate can be obtained in all school types. This certificate attests the skills and abilities acquired at school and entitles the holder—depending on their qualification—to attend various further educational institutions at upper secondary level (for details of the German educational system, Auernheimer, 2005). In our study, about half of the sample came from families with family languages different from German (linguistic migration biography), which is typical for the population of large German cities (Autorengruppe Bildungsberichterstattung, 2016). In sum, we know about many characteristics of the texts' writers. Table 1 shows the composition of the corpus according to the above-mentioned author variables.

Procedure for Report Writers. During corpus generation, Grades 5 to 10 students wrote their reports, next to other tasks, by hand within a session in class. The standardized instruction for this writing assignment was orally presented by the experimenter. Students were given 10 minutes to finish their texts. To ensure comparable amounts of prior knowledge on which the texts

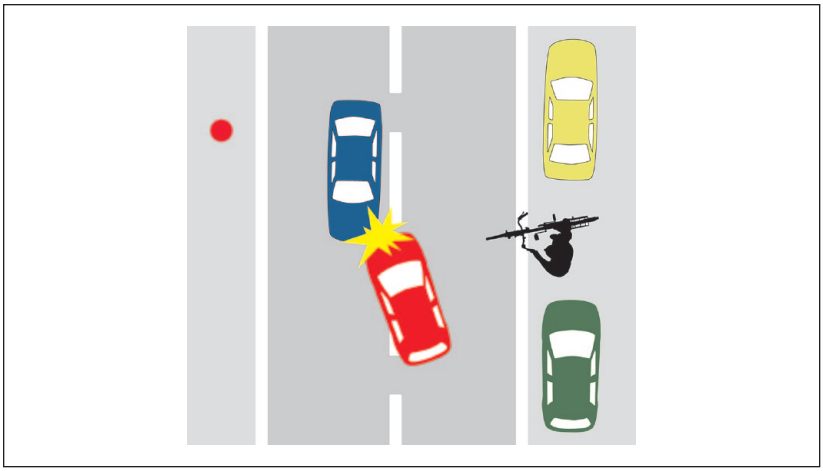

Figure 1. Pictorial stimulus of the writing assignment.
Note. The original picture was colored. The dot represents the observer's position; the car top left is illustrated in blue, bottom left in red, top right in yellow, and bottom right in green. Illustration: Kikkerbillen. Color version of the figure is available online.

are based, the assignment was induced by a pictorial stimulus (Figure 1): writers were asked to imagine that they have witnessed the accident from the perspective indicated by the red point. They were instructed to describe the situation for the police. Altogether, the writing assignment can be considered meaningful, motivating, and context-framed (see Bachmann & Becker-Mrotzek, 2010).

The resulting texts were electronically transcribed and orthographically (but not morpho-syntactically) corrected so that legibility and spelling errors would not bias any assessment of text quality (Greifeneder et al., 2010). These 277 transcribed texts were used in the reconstruction experiment at issue.

Current Study

Participants. University students (N = 277) from Leibniz University Hannover (102 male, 175 female; mean age in years = 23.5, *SD* = 5.5) who successfully participated in a controlled online experiment were recruited from a university-wide database of students interested in experimental participation and confirmed their informed consent. After electronic participation, which took between 5 and 10 minutes, they received a voucher for two candy bars.

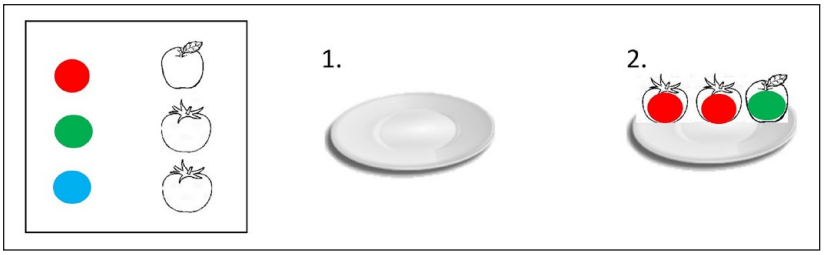

Figure 2. Setup of practice trial.
Note. The original pictures were colored. Three dots on the left: red, green, and blue; three dots on the right: red, red, and green. Color version of the figure is available online.

Procedure of online experiment. The university students received an individual link to a survey page (www.soscisurvey.de) where the instructions and materials were represented. After the login, students were asked to make sure to be undisturbed for the next 5 minutes and to sit in a suitable workplace. If this was not the case, they could start the task later when the environmental requirements were fulfilled. If participation was aborted, or the tasks were otherwise not completed or data were not completely saved, the respective text was reassigned to another participant. Data collection was continued until each of the available 277 pupils' texts had received one valid assessment.

Practice trial. First, the university students completed a practice trial to guarantee that they understood how to manipulate objects and colors in the programmed application. During this trial, students saw a picture on the screen showing two tomatoes and one apple on a plate. The students' task was to reconstruct this image by dragging and dropping the three objects onto the empty plate and marking them with the correct colors (see Figure 2). Only if this practice task was accurately solved, the subsequent main experiment was included for analysis.

Procedure of Accident Reconstruction Task. For the main task, the university students were instructed that they would be shown a text in which a school student described an accident in a witness report to the police. Each of the 277 participants was assigned a different text from the reference study (described above; Grabowski et al., 2014). As explained previously, the task (Figure 1) was based on a scenario where a bicyclist who wanted to cross a road suddenly appeared between two parked cars; a red car was forced to avoid the bicycle, thereby colliding with an oncoming blue car.

The university student's task was to carefully read the assigned text and to reconstruct the described accident with the help of the provided items (cars and bicycles in different orientations, color marks, and the observer's viewpoint; Figure 3). The reconstructing students neither knew the underlying writing instruction nor the associated original pictorial stimulus. According to their best understanding, they dragged and dropped the mentioned objects into the positions described in the text. Additionally, they had to add the respective colors to the objects. Altogether, they arranged the situation that corresponded to their understanding of the accident described in the respective text. Students were given 5 minutes to work on the task. The assigned text remained visible on the computer screen during the entire reconstruction process.

An example of a resulting reconstruction based on one school student's text is shown on the right side of Figure 3, followed by the pupil's text on which this reconstruction was based.

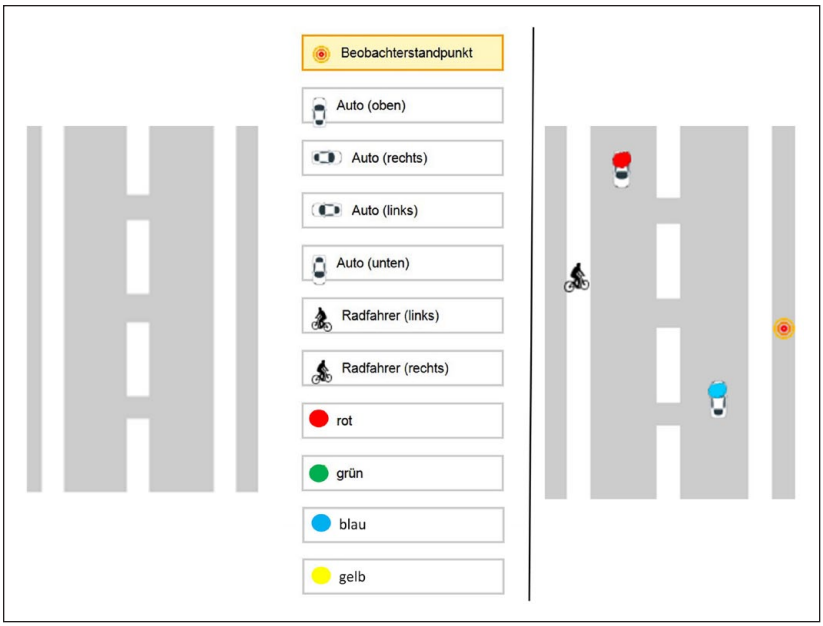

Figure 3. Setup of the main experiment (left) and example of a student's reconstructed solution (right).

Note. The original picture was colored. The provided elements (viewpoint; cars and bicycles in different orientations; colors) could be selected and moved onto the street area: The car above is marked in red, and the car below in blue. Color version of the figure is available online.

The underlying text for this example was as follows (literal interlinear translation is given in brackets [digital transcript loyal to the original handwritten line]):

Das blaue Auto ist normal auf der Straße
[The blue car has normally on the road]
gefahren. Dann ist der Fahrradfahrer gekommen
[driven. Then the bicyclist came]
und wollte die Straße überqueren
[and wanted to cross the road]
und das rote Auto hat sich gefürchtet,
[and the red car was afraid]
dass es gegen den Fahrradfahrer rammt.
[that it hit against the bicyclist.]
Und dabei hat das rote Auto das blaue
[And in doing so the red car has the blue]
Auto nicht bemerkt und ist gegen ihn gerammt.
[car not noticed and has crashed against it.]

Coding and rating of graphic reconstructions. Successful completion of the online experiment resulted in a graphical reconstruction of the accident situation on the basis of a text in which this accident has been described. Files were stored as images. The graphical reconstructions were coded by two raters who both were linguistic student-teachers. They were acquainted with the original picture stimulus (Figure 1) and the writing instruction for the pupils. The reconstructions were coded and rated on three subscales with a total of 16 aspects, plus one holistic overall impression (=17 scores altogether; see Table 2). The aspects were selected according to their relevance for an accurate explanation of the accident situation to the reader.

For each representation, the raters first decided whether or not the six required elements were considered at all (Subscale 1; Aspects 1-6). If the involved objects of the accident, or some of them, were not mentioned or introduced, the proper description of the event would be impaired or even become impossible. On the second subscale, the colors of the four involved cars were rated for correctness (Aspects 7-10). Since there were four schematically pictured cars involved in the situation, their colors were important to establish referential clearness and coherence. The third subscale (Aspects 11-16) related to judgments on whether the physical positions of the respective elements were plausible. The constellation of responsible events that led to the accident can only be understood from the relative spatial relations between the involved objects. Finally, the plausibility of the entire constellation of events leading to the accident, as such, was holistically judged on one

Table 2. Seventeen Aspects for Which the Graphical Reconstructions Were Rated.

Subscale		Required/Obligatory Element of Reconstruction					
		Car 1: veering, red	Car 2: oncoming, blue	Car 3: parking, green	Car 4: parking, yellow	Bicyclist	Observer
1	Does the element exist?	(1)	(2)	(3)	(4)	(5)	(6)
2	Is the color correct?	(7)	(8)	(9)	(10)		
3	Is the position plausible?	(11)	(12)	(13)	(14)	(15)	(16)
	Is the overall situation plausible?			(17)			

item (Aspect 17). We assumed that judgments on the three subscales related, more or less, to simple codings of rather objective facts. Therefore, all codes were binary, that is, 1 (yes) or 0 (no), indicating whether or not the respective characteristic was clearly expressed in the reconstructed picture.

In order to keep the coding format for the raters consistent, rather than changing it after 16 ratings per reconstruction, we decided to rate the holistic overall impression also in the binary format (yes/no). Moreover, by doing so we intended to stress the decision of whether or not a text is *sufficiently* clear to allow for the *sufficiently* exact understanding of the described situation (rather than providing a gradual continuum). Such a binary judgment would also reflect teachers' decisions, for example, in an in-class test, on the basic pass or fail, demarcating the boundary between Grades D and F (United States), or 4 and 5 (Germany), respectively.

Results

Interrater Agreement and Scale Aggregation

Except for 2 of the 17 rating aspects, interrater reliabilities acceptably varied between 84.1% and 99.6% of agreement; the corresponding Cohen's kappa values can be considered substantial or almost perfect (Landis & Koch, 1977; Stemler, 2004). The exact percentages of interrater matches were as follows:

> Subscale 1: Does the element exist? 95.3% (Aspect 1), 93.9% (Aspect 2), 94.2% (Aspect 3), 97.5% (Aspect 4), 99.6% (Aspect 5), and 98.9% (Aspect 6);
> Subscale 2: Is the color correct? 85.2% (Aspect 7), 86.6% (Aspect 8), 97.8% (Aspect 9), and 98.2% (Aspect 10);
> Subscale 3: Is the position plausible? 84.1% (Aspect 12), 96.4% (Aspect 13), 97.8% (Aspect 14), 86.6% (Aspect 15), and 91.0% (Aspect 16).

In contrast, agreement for the position of the veering car (Subscale 3: Aspect 11) and for the overall plausibility of the reconstructed situation (Holistic score: Aspect 17) remained unsatisfactory (72.9% and 69.3% agreement; Cohen's kappa $<.4$). Therefore, these two coding aspects were not further considered, so that Subscale 3 henceforth only comprises five aspects (Aspects 12-16).

We cannot explain why it appears difficult to unanimously evaluate just the position of the veering car (Aspect 11) in the graphical reconstructions. It could be that the static representation is complicated because this car undergoes the most extensive spatial movements during the accident. As regards

the overall impression rated in Aspect 17, the problematic share of rater agreement may reflect the well-known result that holistic judgments on complex structures (either linguistic or pictorial) hinder good reliability; thus, componentially composed scores appear to be more advantageous.

For the remaining 15 rating aspects, the two independent ratings were not unified in cases of disagreement, but averaged, so that finally the ratings had values of 0, 0.5, or 1. These 15 items form an overall scale of acceptable internal consistency (Cronbach's $\alpha = .74$). Therefore, it appeared meaningful to aggregate the 15 averaged items into a single score of text functionality, ranging between 0 and 15. When the aggregate score was computed for both raters separately, interrater correlation was .90, which also indicated high reliability on the aggregate level. Overall, the aggregate score had a mean of 7.3 ($SD = 2.5$), and the median was 7. The worst reconstruction received 2.5 points; the best reconstruction received 14.5 points (out of 15).

The aggregate score (henceforth, functionality score) indicates the similarity between the original situation on which the text was based, and the reconstructed situation based on a reader's understanding of this text. When a text receives a high functionality score, it means that the quality of the text is such that it enables the reader to mentally create a representation of the accident that strongly resembles the original scene on which the writer's report has been based: the text is a functional vehicle to transfer the writer's mental conception of the situation into the reader's mind. For a report text type, a functionality score is a concrete indication of text quality. (Other text genres, as well as other types of reporting writing tasks, may have different functional criteria.)

Plausibility of the Functionality Score

In the reference study from which the text corpus was generated, as well as in similar subsequent studies within our research group (Grabowski et al., 2014, 2018), it turned out that many measures of writing skill, cognitive (e.g., perspective taking) and linguistic (e.g., vocabulary) prerequisites, as well as text-quality measures significantly differed between fifth- and ninth-graders, between educational school type levels, and between students with and without linguistic migration biographies. On average, ninth-graders wrote better (i.e., higher scoring) texts than fifth-graders, pupils from the high-level school type wrote better texts than pupils from the medium- and lower-level school types, and students with German as their only family language wrote better texts than students from families speaking languages different from German. Even if we assume that our functionality score of text quality captures some unique aspects of writers' abilities beyond typical (holistic or

componential) ratings-based text-quality measures (described in next paragraph), the functionality score is still an operationalization of the text-quality construct in our research study. Thus, from the perspective of convergent validity, it can be expected that the systematic dependencies of quality measures described above (on grade, educational level, and family language) would be replicated.

An analysis of variance of the functionality score with the three factors—grade, school type, and linguistic migration biography—from the writers of the respective texts yielded significant main effects of grade, $F(1, 264) = 51.7$, $p < .001$, $\eta_p^2 = .16$; $M_{grade5} = 6.2$, $M_{grade9} = 8.5$; school type, $F(2, 264) = 4.2$, $p < .02$, $\eta_p^2 = .03$; $M_{low} = 7.0$, $M_{medium} = 6.7$, $M_{high} = 8.0$ (post hoc Scheffé test showed two homogeneous subgroups: low + medium vs. high educational level); and linguistic migration biography, $F(1, 264) = 10.6$, $p < .001$, $\eta_p^2 = .04$; $M_{without} = 8.0$, $M_{with} = 6.6$). There were no significant interaction effects. Thus, the quality of reporting texts that enabled readers to reconstruct the described accident situation concordant with the original situation was better when the writers were from ninth grade (as opposed to fifth grade), attended a school of higher educational level (as opposed to medium and low educational levels), and came from families in which only German was spoken (as opposed to families with family languages different from German). This result supports the assumption that the functionality score that we developed as a direct functional measure of text quality relates to the same text-quality construct as other typical measures in which readers rate the quality of texts on a linguistic basis. At the same time, in addition, this result also repeats the typical patterns of advantage and disadvantage received from typical raters (which may not do justice to, e.g., sources of social or linguistic diversity).

The relations to other measures of text quality that were obtained in the study from which the text corpus was taken will be analyzed in the next section, along with specific descriptions of these reference measures.

Correlations With Other Measures of Text Quality

Apparently, measuring the quality of reporting texts via the (mental) reconstructions of their readers is methodologically and empirically more laborious than having raters to assign values to texts after reading. Is it worth it? In other words: does the functionality score capture unique portions of text quality that are not already determined by more typical approaches to text-quality measurement? To answer this question, we compared the functionality score results to three other measures of text quality that were taken for the same texts in the reference study. Subsequently, we will briefly characterize these

alternative measures before we relate them to the functionality score; they are described and justified in greater detail by Grabowski et al. (2014). Note, however, that these comparison scores belong to the reference study. In that reference study, aimed at the prediction of writing competence through several cognitive and linguistic skills of the writers, the text-quality construct was taken for granted. Thus, we did not empirically collect or recalculate the ratings for the present study, but simply (re-)used them for the comparisons with the new functionality score since we knew how the corpus texts from the reference study had been rated with respect to their text quality within the subsequently described approaches.

Text length. After pupils have acquired the basic literacy skills of reading and writing (in the sense of transcription) during primary school, they expand and differentiate their text production skills during secondary grades. Here, text quality is usually strongly correlated with text quantity, that is, the number of words in a text. Particularly for informative texts like reports or instructions, text length can, therefore, be considered a simple quantitative indication of text quality. (For a critical discussion of the relation between text length and text quality in writing assessment, see Fleckenstein et al., 2020.)

Holistic rating. In a first attempt to investigate the possibilities of including writing tasks in German large-scale studies of educational success, researchers of the *Institute for Educational Quality Improvement* at Humboldt-University in Berlin have translated and adapted the *NAEP Holistic Scoring Guide* (Persky et al., 2003). After some explorative studies, the group decided (Schipolowski & Böhme, 2016) to use only five levels in their translated rating scale (as opposed to the six levels defined by Persky et al., 2003). The research team of the reference study (Knopp et al., 2013) received permission to use their version in advance. Technically, the scale for informative texts was used in the current study. (There is a parallel scale for argumentative texts; Schipolowski & Böhme, 2016.)

The implemented text rating scheme is an approach to assessing text quality in a holistic way along a five-level scale (1 to 5; plus level 0 for lowest-quality texts below measurability). Each of the five levels is holistically described in terms of typical characteristics of the to-be-assigned texts with respect to content, structure, syntactical and lexical realization, and linguistic correctness (grammar, spelling, and interpunction). Implicitly, as with most holistic rating scales, it is assumed that the quality of the diverse aspects of a text varies together across the levels. (This assumption may certainly be questionable, e.g., for texts of L2 writers.) The basic intention is to determine the respective level of text quality that best fits the aforementioned characteristics.

Each student's text was independently rated by two student-teachers who were trained with benchmark texts before. The authors report an intraclass correlation between the two raters of .64, and an interrater agreement of 83% when a difference of one level between two ratings is accepted (Grabowski et al., 2014, p. 154). In the face of these moderate degrees of agreement, the two ratings per text were averaged in order to robustify the reliability of the scorings.

Componential rating. The third approach was developed with the aim of creating a robust and efficient measure of text quality. Here, ratings were performed by "naive" linguistic experts, that is, L1 student-teachers without special training. Rather, texts were evaluated on six aspects, or quality components, with dichotomous judgments: overall text quality (high/low), text function (fulfilled/not fulfilled), knowledge necessary for understanding (explicit/implicit), thematic coherence (given/not given), vocabulary (appropriate/not appropriate), and degree of reader-orientedness (high/low). Each text was independently rated by two student-teachers; the two ratings were averaged and aggregated across the six components, resulting in rating scores between 0 and 6. Aggregation appeared justified insofar as the six aspects formed a scale of substantial internal consistency (Cronbach's $\alpha = .90$; Grabowski et al., 2014, p. 155). It was assumed that an aggregated score of six more or less independently assessed aspects would be quite robust with respect to the indicated text quality, while the simple procedure of collecting six yes/no decisions per text did not require high rating expenses. Together, this approach was an efficient approach to score text quality.

In many studies, only one methodological path to the empirical assessment of text quality was chosen, according to the available resources for the rating procedure and the disciplinary traditions of good quality rating. In the reference study, in contrast, three alternative approaches were used in order to broadly cover the aspects that may have contributed to the judgment of how good a text was, as evaluated by raters. With respect to the question whether the functionality score represented, or at least included, a unique rate of what may be considered the quality of a text, this broad empirical coverage appeared to provide a good comparison.

We present three statistical approaches to estimate the degree of uniqueness that was achieved through the empirical determination of the text functionality score, namely, correlation, regression analysis, and factor analysis. First, correlations between the functionality score and the three text-quality measures described above are shown in Table 3.

It happens that the interrelations within the three approaches that use linguistic characteristics of the texts (number of words, or ratings) were

Table 3. Correlations Between the Text Functionality Score and Three Other Measures of Text Quality.

Variable	1	2	3	4
1. Functionality score		.36	.46	.43
2. Text length	.36		.60	.52
3. Holistic rating	.46	.60		.75
4. Componential rating	.43	.52	.75	

Note. All coefficients are statistically significant ($p < .001$; $n = 277$).

generally higher than the correlations between the functionality score and the other quality measures. Thus, the functionality score appeared to be sufficiently related to the more typical indications of text quality, but it may also have contributed some unique aspects that were not covered by the other approaches. Note that the functionality score was empirically derived from the manifest, i.e., non-linguistic reconstructive behavior of readers on the basis of their understanding, while the other approaches bear on judgements of linguistic text characteristics.

In order to further explore the relations between the different measures, we performed a linear regression analysis to show to which degree the functionality score could be predicted by the other three measures. We found that the two rating approaches (holistic and componential), but not text length, significantly predicted the functionality score, $F(3, 273) = 28.1$, $p < .01$, with $R^2_{corrected} = .23$ indicating only a moderate amount of explained variance. Thus, it appears that the direct indication of text functionality captures unique portions of variance that are not bound by the typical rating approaches to text quality.

As a third approach to study the relation of the functionality score to other measures of text quality, we subjected all four quality measures (text length, holistic rating, componential rating, and functionality score) to an exploratory factor analysis (principal components analysis) in order to examine whether the four measures shared the same portions of variance. When the analysis was performed with an eigenvalue >1 as the criterion for extraction, a one-factor solution results that explains for 64.7% of variance. The factor loads on this factor were as follows: text length, .77; holistic rating, .90; componential rating, .86; and functionality score, .67. This factor supports the interpretation of one latent general text-quality source.

However, if we forced factor analysis (principal components analysis) to extract two factors, and requested varimax rotation (with Kaiser normalization), explained variance increases increased to 81.5%. Table 4 shows the

Table 4. Rotated Component Matrix of Two-Factor Solution (Principal Components Analysis).

	Component 1	Component 2
Text length	**.82**	.11
Holistic rating	**.86**	.28
Componential rating	**.83**	.28
Functionality score	.24	**.97**

Note. Coefficients in bold indicate the interpretational simple structure of the factor solution.

factor loads on the extracted components. Now, the first component represents a cluster of the three comparison measures (factor loads > .80), whereas the second component is predominantly fed by the functionality score, with a factor load of .97 and an increase of explained variance by 27.7%.

Altogether, the functionality score is substantially related to other operationalizations of text quality. The functionality score also adds a specific component that enables the reader to evaluate to which degree a text appears to be comprehensible *and* to more or less precisely construct a mental representation of the described situation—which is the immediate communicative goal of a report.

In the final section, we will discuss—beyond the methodological contribution of the approach—how the direct experience of text functionality can be useful in the didactical context of effective writing instruction.

Discussion

The purpose of our study was to develop and implement an alternative methodological approach to the indication of text quality, using a direct functional measure of the reader's understanding. For this study, a score was generated that indicated the similarity between an original accident situation on which each German secondary student's report was based, and the reconstructed situation based on the university reader's understanding of the student's text.

We aimed to address two research questions, the responses to which we are finally able to provide:

1. Does the quality of reporting texts, as measured through the readers' reconstructions of the described accident situation, systematically vary according to student writers' grade, school type, and family language?

This question can be answered affirmatively. The functionality score that we developed as a direct functional measure of text quality repeated the same general patterns as other typical (i.e., holistic or componential) text-quality measures. Thus, the accident reports produced by the students, on the basis of which the readers reconstructed the described situation, were more similar to the original situation if the writers attended ninth grade (as opposed to fifth grade), attended a school with a higher level of education (as opposed to medium and low levels of education), or spoke only German at home (as opposed to family languages other than German). It can, thus, be concluded that the functionality score does not reflect an arbitrary aspect of text effects as it shows the dependencies on grade, educational level, and linguistic migration biography that have been observed with common text-quality ratings. Our method can be considered, therefore, as meaningful in relation to other, typical text-quality measures.

2. How does the functional measure correlate with other approaches to text-quality assessment? Does it simply capture the same aspects, or does it reveal some unique variance of text quality?

We compared the functionality score to three other measures of text-quality that have been used in the context of the study in which the text corpus was generated: text length, a holistic rating, and a componential rating. Overall, data analysis revealed two insights: On the one hand, the functionality score had medium-sized correlations with these three measures. Furthermore, when conducting an exploratory factor analysis, our score loaded at the same text-quality factor when the analysis was performed in the usual grain size. On the other hand, correlations between the functionality score and the other measures were lower than correlations within these three other measures, and in a sharpened factor analysis, functionality bound a unique portion of explained variance. Together, we can assume that the text functionality score covers a distinct facet of text quality that is not—or to a lesser degree—represented by rating approaches to text-quality measurement.

With respect to the immediate evidence from the relation between text quality and its indications, the direct measure of text functionality is unbeatable: it clearly shows whether or not a report is good enough to fulfill its main function, namely, to guide the reader's mental reconstruction and understanding of the reported event as close to the original as possible. In addition, the approach prevented raters from being mainly influenced by linguistic parameters such as vocabulary, grammar, or spelling during their text evaluation. Rather, they only have to assess whether or not the respective obligatory

elements of the accident situation were appropriately mentioned in the text. Therefore, we can assume that our way of assessing text functionality was objective and reliable, which is also indicated by high internal consistencies of the subscales. However, compared to traditional measures of text quality like global ratings, our proceedings were costly—not only time-consuming, but also particularly "participant-consuming." Each text needed a fresh judge who would become unusable for any further text evaluation because they were no longer unbiased with respect to the described situation. So the question arises whether, or to which degree, it is worth approaching text quality in this elaborate functional way. From our point of view, the approach is advantageous, considering that writing itself is a communicative act between writers and readers. With our method we were able to show whether the reader could mentally form a picture of the reported event (here: the accident) and, consequently, whether the text was successful from the reader's point of view. However, the elaborated procedure is only applicable to one particular writing task. In order to examine the functionality of any text, the approach must always be adapted to the task and its underlying situation, which would require much effort and would likely not work for every text genre (and perhaps not even for every kind of reporting texts) with similar objectivity and reliability. The most suitable genres for this kind of procedure are certainly informative texts (rather than argumentative texts), particularly—besides reports—all kinds of instructions (e.g., assembly instructions, experimental protocols, or recipes) where the communicative success (or failure) can be clearly observed.

From a didactical perspective, the approach fits with initial concepts for effective writing instruction by the observation of readers (see e.g., Rijlaarsdam et al., 2006, 2008, 2009). We follow the problem-solving view of writing: the writer is supposed to solve a communicative problem by producing a text that fulfills the communicative needs of the writer and the reader. In our case, the text was successful if the reader mentally reconstructed an accident situation that strongly resembled the original scene. Normally, in classroom writing, students cannot be sure whether their written texts are functionally successful because they never get the chance to put their solution to the test. Instead, they get feedback in form of corrections within the text and written comments by the teacher. But these types of feedback do, at best, indirectly help writers improve their writing in communicative and functional aspects. Moreover, the writers themselves are not able to read their texts like a real reader; they have always prior knowledge, and they seldom recognize the vague passages within their texts. Collaborating with a partner in writing classes can be more effective than only receiving feedback from the teacher on surface-level text features such as spelling or vocabulary.

Functional reader feedback, where partners check each other's text for impact and accuracy, can lead to deeper learning effects for writers.

In educational contexts, studies have already shown that the didactical strategy of observing whether others can make use of one's texts improves students' writing. Schriver (1991, 1992) assessed the influence of reader think-aloud protocols that writers got for their text revisions. Writing novices and experts both showed significant increases in their writing after feedback. So real reader feedback helps writers know more about the communicative needs of readers. Similar results were obtained by Rijlaarsdam et al. (2008, 2009). They showed that reader feedback, reader observation, and role switching between writers and readers can be important supplements to a cognitive, process-orientated perspective on writing instruction. Observation of real readers who actually employ the text for the intended communication purpose (in our case: getting a mental picture of an event) enables writing students to gain feedback for text revision. But students also develop transferable knowledge about readers' needs and behaviors as well as criteria for effective texts of a particular genre. The experience with real readers probably contributes more to the development of audience awareness (Carvalho, 2002) than the traditional practice of learning to write with an imaginary audience in mind. Collaborating with a partner allows writers to focus more on the communicative aspect of writing and aim for the desired effect rather than just fulfilling a checklist of criteria for school performance assessments. As a result, students may also become more willing to take creative risks and experiment with their writing.

According to Rijlaarsdam et al. (2009), the acquisition of skills in a complex domain such as writing always relies on observation and inquiry. They argue for implementing observation as a learning activity in writing education rather than writing a text and revising it afterwards on the basis of correction. They understand writing as an interactive learning activity that stimulates learners' reflection, both as writers and readers. Moreover, this kind of instruction leads to inductive learning of genre and to self-learning, which reduces the teachers' responsibilities. Collecting direct functional measures of text quality can provide a didactical methodological tool in the context of such educational strategies that already involve a component of (self-)assessment.

In conclusion, our approach to direct functional text-quality assessment could be a starting point for classroom intervention and instruction. Based on their experiences with a reconstruction task, students could use similar writing tasks in class and learn to consider the readers' needs and problems with insufficient texts. Afterwards, a repeated functionality test could possibly prove an immediate training effect, and the students themselves might get a

direct sense of success and feel more comfortable and competent in their writing.

However, it is important to note that providing text feedback in this way may not be economical as it requires examination by the reader for each text.

Acknowledgments

We are grateful to Martin Aßmann, who programmed and administered the online experiment.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a grant from the German Ministry of Education and Research (BMBF) to the first author for a research project on subcomponents of writing literacy (grant number 01GJ1208A).

Ethical Approval

Data collection in school for the creation of the analyzed text corpus was approved by the School Authorities of Lower Saxony and North Rhine-Westphalia, respectively. The online experiment was conducted in accordance with existing ethical standards.

ORCID iD

Joachim Grabowski https://orcid.org/0000-0002-0243-721X

References

Auernheimer, G. (2005). The German education system: Dysfunctional for an immigration society. *European Education*, *37*, 75-89. https://doi.org/10.2753/EUE1056-4934370406

Autorengruppe Bildungsberichterstattung. (2016). *Bildung in deutschland 2016. Ein indikatorengestützter Bericht mit einer analyse zu Bildung und migration*. WBV.

Bachmann, T., & Becker-Mrotzek, M. (2010). Schreibaufgaben situieren und profilieren. In T. Pohl & T. Steinhoff (Eds.), *Textformen als Lernformen* (pp. 191-201). Gilles & Francke. http://www.uni-koeln.de/phil-fak/deutsch/sprachdidaktik/koebes/pohl_Steinhoff.pdf

Carvalho, J. B. (2002). Developing audience awareness in writing. *Journal of Research in Reading*, *25*(3), 271-282.

Chan, S., & Yamashita, J. (2022). Integrated writing and its correlates: A meta-analysis. *Assessing Writing*, *54*, 100662. https://doi.org/10.1016/j.asw.2022.100662

Chen, X., & Meurers, D. (2016). *CTAP: A web-based tool supporting automatic complexity analysis*. Apollo—University of Cambridge Repository.

Deno, S. L., Marston, D., & Mirkin, P. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children*, *48*, 368-371.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability* (Research Bulletin RB-61-15). Educational Testing Service.

Ehlich, K. (1984). Zum Textbegriff. In A. Rothkegel & B. Sandig (Eds.), *Text—Textsorten—Semantik* (pp. 9-25). Buske.

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S., & Köller, O. (2020). Is a long essay always a good essay? The effect of text length on writing assessment. *Frontiers in Psychology*, *11*, 562462. https://doi.org/10.3389/fpsyg.2020.562462

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.

Grabowski, J., Mathiebe, M., Hachmeister, S., & Becker-Mrotzek, M. (2018). Teaching perspective taking and coherence generation to improve cross-genre writing skills in secondary grades: A detailed explanation of an intervention. *Journal of Writing Research*, *10*, 331-356. https://doi.org/10.17239/jowr-2018.10.02.06

Grabowski, J., Schmitt, M., & Weinzierl, C. (2010). Second and fourth graders' copying ability: From graphical to linguistic processing. *Journal of Research in Reading*, *33*, 39-53.

Graham, S., Berninger, V. W., Abbott, R. D., Abbott, S. P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology*, *89*(1), 170-182.

Greifeneder, R., Alt, A., Bottenberg, K., Seele, T., Zelt, S., & Wagener, D. (2010). On writing legibly: Processing fluency systematically biases evaluations of handwritten material. *Social Psychological and Personality Science*, *1*, 230-237. https://doi.org/10.1177/1948550610368434

Grobe, C. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. *Research in the Teaching of English*, *15*, 75-85.

Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Erlbaum.

Imbler, A. C., Clark, S. K., Young, T. A., & Feinauer, E. (2023). Teaching second-grade students to write science expository text: Does a holistic or analytic rubric provide more meaningful results? *Assessing Writing*, *55*, 100676. https://doi.org/10.1016/j.asw.2022.100676

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . .

Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Language and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Ke, Z., & Ng, V. (2019). *Automated essay scoring: A survey of the state of the art*. Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19), pp. 6300-6308. https://www.ijcai.org/proceedings/2019/

Knopp, M., Becker-Mrotzek, M., & Grabowski, J. (2013). Diagnose und Förderung von Teilkomponenten der Schreibkompetenz. In A. Redder & S. Weinert (Eds.), *Sprachförderung und Sprachdiagnostik* (pp. 296-315). Waxmann.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Mathiebe, M. (2018). *Wortschatz und Schreibkompetenz. Bildungssprachliche Mittel in Schülertexten der Sekundarstufe I* (Sprachliche Bildung—Studien, Bd. 1). Waxmann.

National Assessment Governing Board. (2010). *Writing framework for the 2011 national assessment of educational progress*. U.S. Department of Education. https://www.nagb.org/publications/frameworks/writing-2011.pdf

Neumann, A. (2012). Advantages and disadvantages of different text coding procedures for research and practice in a school context. In E. van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (Vol. 27, pp. 33-54). Brill.

Nussbaumer, M., & Sieber, P. (1994). Texte analysieren mit dem Zürcher Textanalyseraster. In P. Sieber (Ed.), *Sprachfähigkeiten—besser als ihr Ruf und nötiger denn je!* (pp. 141-186). Sauerländer.

Olinghouse, N. G., & Leaird, J. T. (2009). The relationship between measures of vocabulary and narrative writing quality in second- and fourth-grade students. *Reading and Writing Quarterly*, *22*(5), 545-565.

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing Quarterly*, *26*(1), 45-65.

Persky, H. R., Daane, M. C., & Jin, Y. (2003). *The nation's report card: writing 2002*. U.S. Department of Education/National Center for Education Statistics.

Pohlmann-Rother, S., Schoreit, E., & Kürzinger, A. (2016). Schreibkompetenzen von Erstklässlern quantitativ-empirisch erfassen—Herausforderungen und Zugewinn eines analytisch-kriterialen Vorgehens gegenüber einer holistischen Bewertung. *Journal for Educational Research Online*, *8*(2), 107-135.

Quasthoff, U., & Domenech, M. (2016). Theoriegeleitete Entwicklung und Überprüfung eines Verfahrens zur Erfassung von Textqualität (TexQu) am Beispiel argumentativer Briefe in der Sekundarstufe I. *Didaktik Deutsch*, *41*, 21-43.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*, 2495-2527. https://doi.org/10.1007/s10462-021-10068-2

Rijlaarsdam, G., Braaksma, M., Couzijn, M., Janssen, T., Kieft, M., Raedts, M., Van Steendam, E., Toorenaar, A., & Van den Bergh, H. (2009). The role of readers

in writing development: Writing students bringing their texts to the test. In R. Beard, D. Myhill, J. Riley, & M. Nystrand (Eds.), *The Sage handbook of writing development* (pp. 436-452). Sage.

Rijlaarsdam, G., Braaksma, M., Couzijn, M., Janssen, T., Raedts, M., Van Steendam, E., Toorenaar, A., & Van den Bergh, H. (2008). Observation of peers in learning to write. Practice and research. *Journal of Writing Research*, *1*(1), 53-83.

Rijlaarsdam, G., Couzijn, M., Janssen, T., Braaksma, M., & Kieft, M. (2006). Writing experiment manuals in science education: The impact of writing, genre, and audience. *International Journal of Science Education*, *28*(2-3), 203-233.

Schipolowski, S., & Böhme, K. (2016). Assessment of writing ability in secondary education: Comparison of analytic and holistic scoring systems for use in large-scale assessments. *L1—Educational Studies in Language and Literature*, *16*, 1-22. http://doi.org/10.17239/L1ESLL-2016.16.01.03

Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In E. van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (Vol. 27, pp. 1-22). Brill.

Schriver, K. (1991). Plain language through protocol-aided revision. In E. R. Steinberg (Ed.), *Plain language: Principles and practice* (pp. 148-172). Wayne State University Press.

Schriver, K. (1992). Teaching writers to anticipate readers' needs: What can document designers learn from usability testing? *Utrecht Studies in Language and Communication*, *1*, 141-157.

Shermis, M. D. (2022). Anchoring validity evidence for amtomated essay scoring. *Journal of Educational Measurement*, *59*, 314-337. http://doi.org/10.1111/jedm.12336

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *PARE*, *9*, 4.

Van Steendam, E., Tillema, M., Rijlaarsdam, G., & van den Bergh, H. (Eds.). (2012). *Measuring writing: recent insights into theory, methodology and practice* (Vol. 27). Brill.

Vo, Y., Rockels, H., Welch, C., & Dunbar, S. (2023). Human scoring versus automated scoring for English learners in a statewide evidence-based writing assessment. *Assessing Writing*, *56*, 100719. https://doi.org/10.1016/j.asw.2023.100719

Author Biographies

Joachim Grabowski, full professor of educational psychology at Leibniz University of Hannover, Germany. His main research fields are language production, cognitive processes, and writing.

Moti Mathiebe, postdoctoral researcher at Institute of Psychology, Leibniz University of Hannover, Germany. Her main research fields are vocabulary, academic language, and writing competence.