ASSESSMENT OF WRITING ABILITY IN SECONDARY EDUCATION: COMPARISON OF ANALYTIC AND HOLISTIC SCORING SYSTEMS FOR USE IN LARGE-SCALE ASSESSMENTS

STEFAN SCHIPOLOWSKI & KATRIN BÖHME

Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin

Abstract

Although writing is an important subject of language teaching in secondary education, it is often neglected in large-scale assessments. We report results of a study with 1,365 German high school students in Grade 8 that was conducted in the context of national monitoring of educational achievement. Student responses on seven different persuasive and informative writing tasks were evaluated with two different scoring systems: (i) analytic scoring with 14 dichotomous criteria capturing specific aspects of content, text structure, and language usage, and (ii) holistic scoring based on a comprehensive rating scale similar to the NAEP Holistic Scoring Guide accompanied by semi-holistic scales for content, style (i.e., language usage and organization), and language correctness. We inspected the results of both scoring procedures in terms of inter-rater and intra-rater reliability, dimensionality, and convergence of the scoring results. Attention is also given to the generalizability of the findings across different writing tasks and text genres. The results showed better reliability for the holistic and semi-holistic scales than for most of the analytic criteria. For both scoring systems, content and structural aspects were closely associated whereas language correctness was a clearly distinct dimension. Both scoring systems measured the same latent construct.

Keywords: Writing, analytic scoring, holistic scoring, large-scale assessment, reliability, dimensionality

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, p. 1-22. http://dx.doi.org/10.17239/L1ESLL-2016.16.01.03
Corresponding author: Stefan Schipolowski; Humboldt-Universität zu Berlin, IQB, Unter den Linden 6, 10099 Berlin, Germany; email: stefan.schipolowski@iqb.hu-berlin.de

1. INTRODUCTION

Writing is considered a fundamental skill for success in modern societies (Weigle, 2002; Persky, Daane, & Jin, 2003). Consequently, it is an important aspect of language teaching in secondary education. However, the standardized measurement of these competences is costly and presents researchers with a plethora of methodological challenges, for instance the development of reliable scoring systems. As a result, large-scale assessment programs such as the Programme for International Student Assessment (OECD, 2009) refrain from measuring text production despite its prominent role in school curricula.

In Germany, several large-scale educational assessment programs have been implemented in recent years based on the German *National Educational Standards* that comprise proficiency in the first language in several domains, including L1-writing. However, the objectives of these programs differ: Whereas some aim at educational monitoring on the system level (e.g., Köller, Knigge, & Tesch, 2010; Pant et al., 2013), others focus on detailed feedback for individual schools (e.g., Pant, Gärtner, Harych, Kuhl, & Wendt, 2008). The different aims of the assessments necessitate the development of different essay scoring procedures (e.g., holistic versus analytic coding) that take into account the specific purpose of the respective study. However, the use of different scoring procedures entails the question whether they yield reliable and valid indicators of the same construct, that is, writing ability, and its major facets. In the present study, we examined this research question with regard to two complementary essay scoring systems that have been explicitly developed for use in large-scale assessments in Germany (KMK, 2006). Specifically, we compared a holistic and an analytic scoring approach in terms of (i) intra-rater and inter-rater reliability, (ii) the dimensions of writing ability that can be assessed with each scoring system, and (iii) the question whether the dimensions of writing obtained with different scoring strategies can be regarded as identical constructs despite the differences in scoring methodology (e.g., Is 'language' scored analytically identical to 'language' scored holistically?).

1.1 Holistic and Analytic Essay Scoring

The most common classification of scoring procedures in direct writing assessment is the differentiation between *holistic* and *analytic scoring* (e.g., Cooper, 1977; Knoch, 2009a; Lloyd-Jones, 1977). Some authors additionally consider *primary-trait* and *multiple-trait scoring* (Weigle, 2002). In analytic scoring (e.g., Bryant & Bryant, 2003; Neumann, 2007), a relatively large number of categorical, ordinal or in some cases metric variables (criteria) is used to analyze specific, precisely defined features of a text such as the number of spelling mistakes or the presence of a salutation in a letter. Proponents of analytic scoring promise objective and thus potentially highly reliable scoring criteria as well as a very detailed description of the student's performance in terms of different aspects of writing. Furthermore, analytic scoring offers the opportunity to report profile scores for diagnostic purposes. However, a large number of criteria is needed to describe an essay exhaustively. Scoring such a large number of criteria for each text can be very time-consuming. Also, as there are usually substantial correlations between different scoring criteria, the incremental information gain of a large number of specific text features is questionable (Böhme, Bremerich-Vos, & Robitzsch, 2009). Note that not all analytic scoring systems use very

specific criteria as outlined above but may also refer to larger rhetorical devices such as organizational structure. Often, the use of multiple rating scales to quantify different aspects of writing is already considered analytic scoring (Weigle, 2002). In most analytic scoring systems, the majority of the criteria can be generalized to a class of tasks (e.g., all tasks that focus on the same text genre) or even to different text genres[1]. The fewer scales are used, the more similar analytic scoring becomes to holistic scoring and by rating the quality of rhetorical devices a holistic perspective is included (see below).

According to Cooper (1977, p. 3), "Holistic evaluation of writing is a guided procedure for sorting or ranking written pieces" that "(…) occurs quickly, impressionistically, after the rater has practiced the procedure with other raters." The holistic strategy usually requires the rater to evaluate the quality of a student response on a single rating scale. In other words, the rater considers the text as a whole and quantifies his general impression by integrating content as well as stylistic and language-related aspects in a single score (e.g., Cooper, 1977; Persky et al., 2003). The most basic form of holistic scoring is a general impression marking that is carried out using a rating scale without scale level descriptions. However, to improve reliability holistic scoring is typically based on a scoring guide that provides information about the relevant dimensions of text quality that are to be considered and about the levels of the rating scale (i.e., which aspects should be considered for each level of the scale and the quality of these aspects on each level). Furthermore, examples of student essays (benchmarks) often guide the judgmental process by providing prototypical student responses for the different levels of the rating scale. A prominent example for this strategy is the *Holistic Scoring Guide* of the *National Assessment of Educational Progress* (NAEP; Persky et al., 2003). In the context of large-scale assessments, holistic scoring is time-efficient and provides an indicator of the student's proficiency level that is easier to handle (e.g., for scaling based on IRT models) than a large number of analytic variables. However, scale descriptors are necessarily less specific (i.e., possibly less reliable; Weigle, 2002) and feedback is limited to a single scale score. Therefore, it has been criticized for not providing sufficient diagnostic information for specific interventions (Knoch, 2009b, 2011).

In summary, it can be noted that the different scoring procedures are described and characterized in different ways by different authors and that the transitions between these approaches are fluid. One may benefit from the following model conception: We can classify scoring procedures and the corresponding scales along two separate (but partly correlated) axes. The first axis relates to the object of assessment and answers the question whether elements of a text are looked at or the text as a whole. One could reduce this idea to the question: *What is judged*? The second axis describes the final result of scoring: Does the judgment of a text or text features result in a single comprehensive score or will there be multiple scores? One could reduce this aspect to the question: *What is the final result*? The two axes can be represented in a fourfold table (see Table 1).

A more recent development that has received considerable attention especially in the United States is the use of computer-automated scoring systems (Williamson, Xi, & Breyer, 2012; Shermis & Burstein, 2013). However, despite compelling evidence for the validity

[1] *Sometimes a distinction is made between different rhetorical modes (e.g., description, narration, or argument) and different genres (e.g., letter, essay, or manual) (e. g., Rijlaarsdam et al., 2012) of the writing tasks. We use the term genre in a broader sense to refer to any of these differences.*

of the resulting scores (e. g., Attali & Burstein, 2006), automated essay scoring is limited when the aim is to evaluate more specific aspects of a student response such as language usage or the organization of content (Attali, Lewis, & Steier, 2013). Furthermore, automated scoring requires the development or acquisition of specialized software which may not be feasible depending on the study at hand. Finally, scores from automated systems are still not widely accepted by students and teachers (Lenhard, Baier, Hoffmann, & Schneider, 2007). For these reasons we focus on human scoring in the present article.

Table 1. Fourfold Scheme of Text Scoring

	Object of assessment	
Final result of scoring	*Text as a whole*	*Aspects of a text*
Single comprehensive score	Holistic scale	Sum score
Multiple scores	"Semi-holistic" scales	Analytic scales

1.2 Empirical Questions in Essay Scoring

The evaluation of a large quantity of student responses based on two different scoring approaches in the present study provides insights into crucial questions concerning essay scoring. An important concern is the reliability of the scores as the students' written responses are judged by raters (Bachman & Palmer, 1996; Huot, 1990a). Human raters often diverge in their evaluations of the same text and sometimes do not provide consistent ratings at different points in time (Huot, 1990b; Schoonen, 2005). Hence, analyses of rater inconsistency and score reliability are a prerequisite for valid ability estimates (Congdon & McQueen, 2000). In many cases, rater trainings are implemented to maximize both agreement between different raters (inter-rater agreement) and consistent rating behavior for each individual rater (intra-rater reliability) (Barrett, 2001; Weigle, 1998). However, the effect of rater training on score reliability has been found to be relatively small in some but not all cases (Eckes, 2008; Shohamy, Gordon, & Kraemer, 1992). It is an important empirical question to what extent score reliability depends on the implemented scoring procedure (Shohamy et al., 1992).

Another important issue is the dimensionality of the underlying cognitive ability that is captured with the ratings. Most scoring systems consider aspects of content, text structure or style, and language correctness. It is an empirical question whether this distinction is reflected in the actual ratings and how strongly these aspects are interrelated. Furthermore, as it is desirable to establish a common writing ability scale covering different topics, text genres, and writing purposes, it is essential to show structural invariance of the required ability across tasks and text types (Cumming et al., 2005; see also Tillema, van den Bergh, Rijlaarsdam, & Sanders, 2012). Both the structure of writing ability and its invariance across tasks can be studied separately for each scoring system. It is, however, equally important to examine whether invariance holds across different scoring approaches (Böhme et al., 2009; Goulden, 1994; Slater & Boulet, 2001). The latter pertains to the question whether the same ability dimensions or constructs are captured independently

of the rating procedure. In the following, we briefly review empirical findings regarding these issues.

Bae and Bachman (2010) investigated the validity of four factors of writing ability (grammar, content, spelling, and text length) in English based on elementary school students' responses to letter and story-writing tasks. The student texts were scored analytically and the respective ratings were analyzed with confirmatory factor analysis (CFA) using a model with correlated factors and correlated uniquenesses (i.e., freely estimated correlations between residuals associated with the same task). The authors found very high correlations between latent variables for content and grammar (ρ = .89), content and text length (ρ = .88), and grammar and text length (ρ = .80). Slightly lower correlations were identified between the factors representing content and spelling (ρ = .68), and grammar and spelling (ρ = .70). The lowest correlation was found between spelling and text length (ρ = .53). The authors could show that a higher-order model with one factor on the highest level (second order factor) and four factors on the level below (first order factors) was preferable to a single factor model.

In the context of the German *National Educational Assessment* for elementary school, writing ability of 3rd and 4th graders was studied exploring similar questions as in the present study (Böhme et al., 2009). Comparing the reliability of holistic and analytic scoring for a narrative writing task, satisfactory inter-rater agreement was found for both scoring systems. With regard to the structure of writing ability, the pattern of intercorrelations among three latent variables based on the semi-holistic scales showed that content and style were strongly related (ρ = .88) whereas the correlations of language correctness with content (ρ = .48) and style (ρ = .65) were both considerably lower. Likewise, a global impression rating was very strongly associated with the two semi-holistic scales for content (ρ = .94) and style (ρ = .95) whereas the correlation with language correctness was substantially lower (ρ = .67). This pattern of results indicates at least partially distinguishable dimensions of writing ability. The convergence of the rating systems was strong; e.g., the correlation between latent variables representing content based on the respective scoring strategy ranged from .83 to .87. The authors concluded that both evaluation procedures constitute different methodical approaches to measuring the same construct (Böhme et al., 2009, p. 321f).

Similar findings on the structure of writing ability have been reported for secondary education. Using data from large-scale assessments in 9th and 11th grade, Neumann (2007) analyzed the dimensionality of writing ability for four writing tasks using confirmatory factor analysis. A comparison of different models revealed that a distinction between 'language system' on the one hand and 'semantics/pragmatics' on the other hand was supported best by the data. In Neumann's (2007) study, 'language system' was mainly comprised of variables indicating language correctness whereas 'semantics/pragmatics' referred to aspects of content and structure. Both latent factors constituted empirically distinguishable, but substantially correlated constructs (ρ = .74). However, the analyses were based on a single text genre (i.e., composing a letter).

1.3 The Present Study

Depending on the specific aims of large-scale educational monitoring programs, writing assessments may take place in different contexts and for different purposes which require

different assessment and scoring strategies (Alderson, 2006; Neumann, 2012). Whereas programs implemented to provide feedback on the level of classes or individuals (i.e., information that is useful for teachers to improve writing instruction) require extensive scoring of various aspects of a student response, programs focusing on the evaluation of competences on the system level (i.e., reporting means for different states) are better served by a faster, less expensive, and less detailed scoring approach such as a holistic global impression rating. However, independent of the specific context and scoring, writing assessments should provide reliable and valid indicators of the same construct, that is, writing ability, and its major dimensions. Against this background, our research goal in the present study was to examine whether two different scoring systems based on analytical criteria and holistic rating scales, respectively, lead to comparable results in terms of reliability and validity. This research goal was pursued by examining three intertwined research questions: First, a relevant comparison and starting point was whether the judgments of the raters were reliable for each scoring system in terms of intra-rater and inter-rater agreement (research question one). Second, we investigated the validity of the scores of each scoring system by analyzing their dimensionality. Specifically, we examined whether the judgments of the raters reflected the major dimensions of writing such as content, structure, and language correctness (research question two); these analyses were conducted separately for each scoring system. The third research question was also relevant to establish the validity of the scoring results: By merging the results from both scoring procedures, we investigated whether both scoring systems provide measures of the same ability. Specifically, we tested whether the dimensions of writing ability established with analytic scoring were not different from the dimensions of writing represented by holistic scores. Since task and genre effects have been reported to be very prominent in writing assessment (Bouwer, Béguin, Sanders, & van den Bergh, 2015; Schoonen, 2012), all analyses were conducted separately for each task to identify possible task and genre effects for each scoring system.

Taken together, the analyses shed light on the role of the coding strategy for the precision and interpretation of writing scores. Although the focus of the present study is on large-scale assessment, knowledge about the importance of the scoring system used for evaluating writing is also relevant for practitioners who have to decide on how to score written student responses.

2. METHOD

2.1 Participants

1,365 high school students in 8th grade from 55 schools in eight German federal states participated in the writing assessment. All common school types of the German secondary education system were included in the sample (for details on the German education system, see e.g., Auernheimer, 2005; KMK, 2014). The mean age of the students was 14.66 years (*SD* = 0.65 years), 47.1 percent of the students were female and 22.1 percent of the participants indicated that the first language they had learned in their family was not German. This percentage includes bilinguals and is consistent with the percentage of stu-

dents from families with an immigration background reported in recent large-scale assessments in Germany (e.g., Stanat & Christensen, 2006).

2.2 Measures and Design

The study included seven different persuasive and informative writing tasks developed by a team of experts on German language for use in national educational assessment programs. Three informative tasks required students to write a report (e.g., for a newspaper) based on given pieces of information or to write a description (e. g., of a stage setting). For instance, in the task "Newspaper Report" students were given a picture of a reporter's notepad containing headwords on a story the reporter wanted to write for the next issue of the newspaper. The students then had to use the given facts to write a short newspaper article. Four persuasive tasks presented a controversial standpoint or issue and required the students to discuss it in written form. For instance, in the task "Letter to the Editor" students first had to read a short letter printed in the local newspaper. In this letter, an elderly woman complained about rude behavior of juveniles on the bus (e.g., not offering her a seat) as proof for a general decline in manners. Afterwards, the students were asked to write a letter to the editor on their own in response to the elderly woman's point of view. They were free to either agree or disagree with the woman's standpoint or to take an intermediary position, but they were required to provide convincing arguments for their own position.

To avoid fatigue effects, not all students were administered all tasks. More specifically, a complex rotation design (*balanced incomplete block design*; e.g., Frey, Hartig, & Rupp, 2009) with eight different booklets was used. The number of available cases for each task varied between 320 and 356 (*Md* = 350)[2].

2.3 Procedure

The writing assessment was part of a larger study with additional tasks on other language domains based on the German *National Educational Standards*. All students first completed a 40 minute test of reading comprehension followed by a short break of five minutes. Afterwards, each student was asked to respond to two different writing tasks. For each task, 20 minutes of working time were allotted. All tests were conducted as group tests in the usual classroom setting using paper and pencil. Students were not allowed to use any additional resources such as dictionaries.

2.4 Scoring

Scoring systems. Two different scoring strategies were applied to evaluate the quality of the students' texts: An analytic strategy with 14 dichotomous variables coding specific aspects of content, text structure, and language usage and a holistic strategy using a global impression rating accompanied by three semi-holistic scales.

The analytic scoring criteria are given in Table 2. The decision to use a limited number of dichotomous criteria was made to provide a time-effective scoring procedure that can

[2] *The median (Md) is reported because it is less sensitive to outliers than the mean.*

be easily implemented in the context of large-scale assessments with thousands of participants. Language criteria were identical for all tasks. Structural criteria were identical for a given text genre (i.e., informative versus persuasive writing). Scoring instructions for content criteria were specific for each task.

The holistic global impression rating was a translation and adaptation of the NAEP *Holistic Scoring Guide* (Persky et al., 2003). Each level of the global scale describes the quality of the student text in terms of content, style, and language correctness. Several exploratory studies conducted during the adaptation of the scale led to the decision to use only five substantial rating scale levels instead of the six levels defined by Persky et al. (2003). Using only five levels instead of six reduced the probability of an extremely low percentage of student responses on the lowest scale level. Since the global scale resulted in only one score for each student response without the possibility to distinguish between different aspects of writing proficiency, three additional semi-holistic rating scales were devised to complement the general impression with regard to content, style, and language correctness. These scales were 'semi-holistic' in the sense that each scale required the raters to consider the student text as a whole, but only regarding a specific dimension of writing. In the following, we consider the semi-holistic scales and the global impression taken together as a holistic scoring system because all four scales take the same position on the "Object of assessment"-axis ("Text as a whole"; see Table 1). Due to the more specific focus of the semi-holistic scales on a single dimension, only four different levels were specified. For all levels of each (semi-)holistic scale, original student responses were provided as benchmarks during rating to illustrate the scale levels. The benchmarks were selected by experts on language assessment and writing based on earlier expert ratings of a random selection of student responses.

Table 2. Analytic Scoring Criteria

Language criteria	Structural criteria	Content criteria
Orthography	Abidance by text genre	Task comprehension
Grammar	Reference and recipient orientation	Main content: most important information/ arguments
Punctuation	Coherence and cohesion	Additional information/arguments
Vocabulary	Text structure (e.g., paragraphs)	Richness of the information/argumentation
Syntax		Specific quality feature [a]

Note. Each criterion was scored as either "reached" (score 1) or "not reached" (score 0). [a] *Depending on the task, "specific quality feature" could for instance refer to the presence of an adequate headline or the explicit anticipation of a possible counter-argument.*

Raters and scoring procedure. All texts were scored according to both strategies by two groups of seven raters each. That is, one group was scoring the texts with the holistic and semi-holistic scales whereas the other rater group scored analytically. Raters were randomly assigned to one of the groups. (Note that two raters were part of both groups, which is, however, not relevant for the present study.) In each rater group, every rater scored a random selection of student responses for each of the seven writing tasks. For analytic scoring, a given student response was scored according to all scoring criteria by

the same rater before moving to the next student text. For (semi-)holistic scoring, the scales were applied in separate scoring rounds (i.e., at first all student responses were scored with the global impression scale; in the next round, the texts were presented in a new randomized order and scored with the content scale, and so on). Raters were university students studying for a teaching degree or enrolled in educational science. The majority of raters had prior experience in scoring student responses collected in large-scale assessments.

Prior to scoring, all raters received intensive training. An important part of the training was the evaluation of a large number of student texts (between 40 and 60 for each task, depending on preliminary reliability analyses during training) by all raters. Critical student texts (i.e., texts with low inter-rater agreement) were subsequently discussed in detail under the instruction of an expert in order to establish a common understanding of the scoring criteria and scales. Altogether, there were approx. 80 hours of training for each rater and scoring system. The two rater groups were instructed by different trainers.

To evaluate inter-rater agreement and reliability, for each scoring system between 180 and 210 student texts for each writing task were scored by two different raters from the same rater group. To maximize the generalizability of the results, these texts were randomly selected from the total of available texts. Likewise, the allocation of the individual raters to student responses and the points in time these texts were rated were also randomized. Furthermore, a random selection of 30 texts per task was scored by each rater at the beginning and again at the end of the scoring period to allow for the evaluation of intra-rater agreement. Note that raters were not made aware of any details about the rater design but only instructed that some student responses may appear more than once during scoring.

2.5 Statistical Analyses

To estimate inter-rater and intra-rater agreement for the (semi-)holistic scales, we calculated intraclass correlations (ICC; Wirtz & Caspar, 2002). The ICC coefficient can be interpreted as the proportion of trait variance in the ratings, that is, variance attributable to actual interindividual differences in writing ability (as opposed to variance due to a lack of consensus between raters). In addition, we calculated the percentage of cases with absolute agreement and the Spearman rank correlation coefficient. The latter is suitable for ordinal data and a measure of reliability, that is, leniency and severity effects are not taken into account. For the analytic criteria, the percentage of cases with absolute agreement and Cohen's (1960) kappa coefficient are reported. Kappa involves a correction for agreement by chance. Note that for dichotomous criteria, the magnitude of kappa is comparable to the ICC coefficient (Wirtz & Caspar, 2002). There is no consensus in the literature concerning acceptable values for reliability and agreement coefficients. Following the recommendations by Fleiss, Levin, and Paik (2003), we assume that kappa values between .40 and .50 are the minimum for useful scoring criteria. The same rule is used for ICC coefficients as kappa and ICC values are comparable for the data presented here.

Analyses of the dimensionality of the measured constructs and their interrelations were accomplished with confirmatory factor analysis (CFA) using the software Mplus and the WLSMV estimator which was developed for the analysis of categorical data such as binary variables (e.g., criterion given versus not given) and rating scales (Muthén &

Muthén, 1998-2007). CFA is a technique that allows for the estimation of latent (i.e., error-free) variables (factors) and their interrelations based on observed indicators (e.g., ratings of text quality) and can be used to investigate the internal structure of a given set of variables. For the holistic scales, CFA relied only on those texts which had been evaluated by two different raters. The specific combination of the two raters scoring the same student response was randomly determined and therefore differed for each case. In the analyses, the specific rater combination was ignored; instead, the two judgements available for each respective case were used to define two structurally equivalent "virtual raters" or "pseudo raters" randomly drawn from a rater population (Böhme et al., 2009). The randomized assignment of student responses to raters ensured that the results were not biased by rater effects such as leniency or severity effects or differences in reliability between individual raters. Because they are assumed to be interchangeable, pseudo raters were always modeled with equal intercepts, variances, and factor loadings. Latent factors for the different aspects of writing ability (e.g., content) were established using the ratings of the two pseudo raters as indicators[3]. For the analytic criteria, all texts including those that had been evaluated by only one rater were used in the analyses. Ratings of the dichotomous criteria served as indicators for latent factors. Model fit (i.e., information on how well the model describes the data) was evaluated based on commonly used statistics, namely the chi-square value (χ^2) and degrees of freedom (*df*)[4] as well as the comparative fit index (CFI), root mean square error of approximation (RMSEA) and weighted root mean square residual (WRMR). According to Yu (2002), models with categorical data and good model fit are characterized by the following values: CFI ≥ .96, RMSEA ≤ .05, and WRMR ≤ .95.

3. RESULTS

Between 320 and 356 student responses were available for each task. However, not all responses could be scored due to missingness and nonsensical responses implying an intentional disregard of task instructions. Depending on the task, between 2.2% and 14.4% of the responses were excluded (*Md* = 7.5%), leaving 303 to 337 evaluable student texts per task for the following analyses. On average, the students wrote 76 to 100 words for each task.

---

[3] As pointed out by an anonymous reviewer, the evaluation of text quality has to be distinguished from the assessment of writing ability. Scoring systems provide ratings of text quality; such ratings can, however, be viewed as indicators of writing ability (Rijlaarsdam et al., 2010). It has been argued that a reliable assessment of writing ability for individual feedback requires more than a single writing task (see e.g., Bouwer et al., 2015). In the present paper, text quality ratings are regarded as manifest indicators of (latent) writing ability and analyses aim at conclusions on the group or population level. Analyses were conducted separately for each task since we intended to investigate task-specificity and because the individual student was administered only two out of seven different writing tasks.

[4] Note that in WLSMV estimation, the degrees of freedom are estimated rather than computed; therefore, neither df nor χ^2 can be interpreted in the same way as in models with maximum likelihood (ML) estimation.

In the following, results are presented concerning research questions one (reliability), two (dimensionality of the writing scores), and three (relationships between the dimensions of writing scored analytically versus holistically).

3.1 Inter-Rater and Intra-Rater Agreement

Holistic and semi-holistic scales. For the global impression rating, depending on the task, ICC coefficients for inter-rater agreement varied between .61 and .81 (*Md* = .69). Rank correlations were almost identical to the ICC coefficients. Perfect agreement between raters was found for 48% to 63% of the cases (*Md* = 57%). However, the relatively low percentage agreement is a consequence of the larger number of categories on the global impression scale (in comparison to dichotomous criteria); if a deviance of +/- 1 level on the scale was allowed, agreement was at 95% to 99% (*Md* = 97%). Reliability and agreement results for the semi-holistic scales were similar to the results for the global scale with inter-rater ICC coefficients of .57 to .74 for the content scale (*Md* = .71), .55 to .78 for the style scale (*Md* = .62) and .59 to .75 for the language correctness scale (*Md* = .66). In general, there were no substantial differences between the text genres. Intra-rater agreement was also very similar for all scales with ICC values between .82 and .89 (*Md* = .84) depending on the task.

Analytic criteria. Across all tasks, kappa coefficients for inter-rater agreement varied between .28 and .61 in the language domain (*Md* = .44), .28 and .80 in the structure domain (*Md* = .49), and .15 and .66 in the content domain (*Md* = .46). Variability was equally high for percentage agreement (64% to 92% across all criteria; *Md* = 83%). Hence, some analytic criteria did not consistently fulfill basic requirements for inter-rater agreement. Most notably, ‘grammar’, ‘vocabulary’, ‘text structure’, and ‘specific quality feature’ had kappa values close to or below .30 for several tasks. ‘Vocabulary’ proved to be especially problematic as it consistently showed very low reliability on all tasks and was therefore excluded from further analyses. For the same reason, ‘specific quality feature’ was excluded from further analyses of persuasive writing tasks (but not for informative tasks); in this case, the lack of reliability was due to an extremely skewed score distribution. Intra-rater reliability was lower for analytic criteria than for the (semi-)holistic scales: Depending on the criterion, kappa values ranged between .48 and .79 (*Md* = .71). Inter-rater reliability analyses were also conducted for composite scores (i.e., sum scores) based on the analytic criteria: Considering all tasks, the median ICC coefficient for a total score (unweighted sum of all 14 criteria) was .73. For the domain scores, the respective ICC coefficients were .67 (language), .63 (structure), and .60 (content).

Summary regarding research question one. The holistic global scale, all three semi-holistic scales and the majority of the analytic criteria showed acceptable reliability as a prerequisite for a valid assessment of writing in the context of large-scale assessments for the purpose of system monitoring (see Table 3 for a summary of the inter-rater results). With holistic scoring, about 60% to 70% of the variance in the ratings represented trait variance (i.e., individual differences in writing ability). Reliability was lowest for the style aspect, although the difference to the other scales was small. For the analytic criteria, reliability coefficients showed higher variance and were more task-specific. Here, the lowest reliability was found for vocabulary. The average kappa value of .46 across all tasks and criteria indicates that slightly less than half of the variance in the analytic ratings reflects

differences in writing ability. In other words, about half of the variance in the analytic ratings has to be considered measurement error, which is at the lower end of what has been proposed as acceptable agreement between raters (e.g., Fleiss et al., 2003). On the other hand, reliability of composite scores based on the analytic criteria was on par with the holistic and semi-holistic scales.

Table 3. Summary of Inter-Rater Reliability Results

Scale/aspect	Holistic scoring [b]	Analytic scoring	
		Criteria [a]	Sum score [b, c]
Global/total	.69	.46	.73
content	.71	.46	.60
style/structure	.62	.49	.63
language	.66	.44	.67

Note. [a] Median of the kappa values of the respective criteria, across all tasks. [b] ICC coefficients. [c] Sum scores based on all respective criteria (see Table 2); for homogeneity analyses, see measurement models in Dimensionality of Writing Ability.

3.2 Dimensionality of Writing Ability

Holistic and semi-holistic scales. Correlations between latent variables based on the global impression rating and content, style, and language correctness as evaluated with the respective holistic and semi-holistic scales are given in Table 4 for two informative and two persuasive tasks as examples. Note that these correlations reflect the associations between latent variables and are thus corrected for measurement error which is conceptualized as disagreement between the two pseudo raters for the same student text and rating scale. For all four tasks, very strong relationships existed between the global impression and the latent variables reflecting content and style, respectively. Language correctness was also substantially correlated with the global impression, but considerably less than content and style variables. Among the factors based on the semi-holistic scales, content and style were very closely associated. In one case (persuasive task 1), this correlation was not different from unity ($\Delta\chi^2(1, n = 166) = 2.8$, $p = .097$) indicating that content and style were not empirically distinguishable for this task. Again, language correctness set itself apart from both content and style while maintaining a strong relationship to both aspects, especially with style. As indicated by the results in Table 4, the findings were relatively homogeneous across text genres and tasks.

Table 4. Correlations Between Latent Variables Based on the Global Impression Rating and the Semi-Holistic Scales for Content, Style, and Language Correctness for Four Writing Tasks

Scale	1. Global impression	2. Content	3. Style
Persuasive task 1			
1. Global impression	-		
2. Content	.97	-	
3. Style	.94	.95	-
4. Language correctness	.65	.62	.80
Persuasive task 2			
1. Global impression	-		
2. Content	.96	-	
3. Style	.97	.85	-
4. Language correctness	.66	.65	.77
Informative task 1			
1. Global impression	-		
2. Content	.98	-	
3. Style	.93	.87	-
4. Language correctness	.74	.63	.87
Informative task 2			
1. Global impression	-		
2. Content	.94	-	
3. Style	.92	.79	-
4. Language correctness	.76	.64	.82

Note. N varies between 166 and 184 depending on the task.

Analytic criteria. Two models were compared for each of the four tasks (i.e., the same four tasks considered in the previous paragraph). In the first model, three dimensions (i.e., latent variables/factors) representing language, structure, and content were defined. Hence, this model (3-factor model) reflected the theoretical classification of the criteria depicted in Table 2. The 3-factor model was compared to a 2-factor model based on the distinction made by Neumann (2007) who differentiated language aspects on the one hand from structural and content aspects on the other hand. Specifically, for the 2-factor model the language factor was retained as a factor called 'language system' whereas all other criteria contributed to a factor designated 'semantics/pragmatics'. In other words, in line with Neumann's (2007) results, the 2-factor model was based on the assumption that content and structure are indicators of the same dimension of writing ability. Thus, a comparison of both models in terms of model fit (i.e., Which model describes the data best?) allowed for a test of the empirical distinguishability of structure and content.

Note that the criterion "task comprehension" had to be excluded from all analyses because of a pre-defined dependency with the other content criteria which violated modeling assumptions. Specifically, students who did not evince task comprehension automatically received "not reached" scores on the other content criteria, creating an artificial correlation of those criteria with 'task comprehension'.

For three of the four tasks, the initial model specification in accordance with Table 2 (3-factor model) showed acceptable model fit. However, depending on the task, one or two residual correlations were substantially higher than 0 and had to be freely estimated to achieve acceptable overall model fit. This applied to 'orthography' and 'punctuation' in

all cases. (Note that a significant residual correlation between two variables means that they share common variance above and beyond the variance explained by the latent variable.) For one of the four tasks (informative task 1), acceptable model fit could only be established by dropping two structural criteria, 'coherence and cohesion' and 'text structure'. This means that these criteria were not psychometrically adequate for the measurement of structure in this task.

A comparison of the 3-factor model with the model specification according to Neumann (2007; 2-factor model) revealed significantly lower model fit for the 2-factor solution for two of the four tasks: for persuasive task 1, $\Delta\chi^2(2, n = 316) = 26.4$, $p < .001$; for informative task 2, $\Delta\chi^2(2, n = 311) = 11.9$, $p = .003$. Hence, the initial specification with three factors was retained for these two tasks. For the other two tasks, there was no significant difference in model fit between the 2-factor specification and the 3-factor solution: for persuasive task 2, $\Delta\chi^2(2, n = 321) = 5.0$, $p = .083$; for informative task 1, $\Delta\chi^2(2, n = 317) = 1.3$, $p = .535$. Hence, for these tasks, the 2-factor model provided a more parsimonious description of the data and was thus preferable. This means that for two of the four tasks, there was no empirical evidence supporting the distinction between content and structure. Model preference was not associated with a specific text genre. See Table 5 for fit statistics for all four tasks and both models.

The correlations between the latent factors for the preferable solutions are given in Table 6. In comparison to the results for the holistic evaluation, the task dependency of the results was more pronounced for the analytic criteria. Regarding the two tasks where the 2-factor solution was superior, a substantial correlation between language and 'semantics/pragmatics' was found. In both cases, this correlation was different from unity and for one task (persuasive task 2) of similar magnitude as reported by Neumann (2007). The lower correlation for the second task (informative task 1) may be due to the underrepresentation of structural criteria in the factor 'semantics/pragmatics' (see note given for Table 5). For the other two tasks, the model comparison supported a distinction between content, structure, and language aspects. Here, the highest correlations were found between language and structure. The correlations between content and language were considerably lower, but still substantial. The association between content and structure was inconsistent between the tasks.

Table 5. Model Fit Indices of Competing Measurement Models for the Four Writing Tasks

Task	Model	*N*	Model fit					
			χ^2	*df*	*p*	CFI	RMSEA	WRMR
Persuasive task 1	3F[a]	316	34.6	27	.150	.987	.030	0.741
	2F	316	60.5	28	.000	.946	.061	0.995
Persuasive task 2	3F	321	42.0	25	.018	.970	.046	0.830
	2F[a]	321	45.2	26	.011	.966	.048	0.876
Informative task 1	3F	317	41.4	20	.003	.967	.058	0.883
	2F[a]	317	41.2	21	.005	.969	.055	0.896
Informative task 2	3F[a]	311	49.7	31	.018	.962	.044	0.869
	2F	311	60.9	32	.002	.941	.054	0.973

Note. For informative task 1, two structural criteria had to be dropped from the analyses to achieve acceptable model fit (see text for details). For all tasks, the correlation between the residuals of "orthography" and "punctuation" was freely estimated. Additional residual correlations were allowed for persuasive task 2 (between "abidance by text genre" and "reference and recipient orientation") and informative task 2 (between "richness of the information" and "specific quality feature"). 3F = 3-factor model (see text), 2F = 2-factor model (see text), CFI = Comparative fit index, RMSEA = Root mean square error of approximation, WRMR = Weighted root mean square residual.

[a] Preferred model (see text).

Table 6. Correlations Between Latent Variables Derived From the Analytic Evaluation of Four Different Writing Tasks

Three-factor model	1. Language	2. Structure
Persuasive task 1		
1. Language	-	
2. Structure	.83	-
3. Content	.54	.73
Informative task 2		
1. Language	-	
2. Structure	.91	-
3. Content	.57	.53
Two-factor model	2. Language system	
Persuasive task 2		
1. Semantics/pragmatics	.73	
Informative task 1		
1. Semantics/pragmatics	.48	

Note. N varies between 311 and 321 depending on the task. For additional information on model specifications, see text and Table 5.

Summary regarding research question two. Dimensionality analyses for both evaluation procedures revealed that language and content were relatively distinct aspects which was also true—with limitations—for the distinction between language and stylistic/structural aspects. Analyses based on the semi-holistic scales showed that a separation of style and content was difficult from an empirical perspective. Also, both content and style ratings

were very strongly correlated with the global impression. With analytic scoring, results were relatively strongly dependent on the task but not on the text genre.

3.3 Relationships Between Latent Variables Based on Different Scoring Systems

To verify whether both scoring systems measured the same or different constructs, the measurement models established in the previous section were combined in a structural model. More specifically, for each task, the structural model contained latent (i.e., error-free) variables representing the ratings on the semi-holistic scales on the one hand and latent factors representing the prevalent dimensions of the analytic ratings on the other hand. Hence, the structural model enabled the estimation of the relationships between the latent variables derived from both scoring systems. Latent variables were established using the ratings of the pseudo raters as indicators (see above). For the analytic factors, sum scores were computed separately for both pseudo raters according to the results of the dimensionality analyses; the sum scores were then used as indicators for the latent variables. The estimation was carried out separately for the four tasks analyzed in the previous section (i.e., two informative and two persuasive tasks).

Results for informative task 1 are shown in Figure 1. The results indicate a very strong correlation between the holistic content factor and the analytic factor 'semantics/pragmatics' on the one hand and an equally high association between the two language factors on the other hand. Fixing the correlations between the respective latent variables to 1 entailed a non-significant change in model fit, $\Delta\chi^2(5, n = 178) = 3.8$, $p = .583$. In other words, the semi-holistic content scale and the analytic criteria that defined the 'semantics/pragmatics' factor measured identical constructs; the same applied for the measurement of language aspects in both scoring procedures. Furthermore, style as measured by the semi-holistic scale was highly associated with both analytic factors. In accordance with the results from the dimensionality analyses, both language factors could be empirically distinguished from factors representing content.

Results for the other three tasks are summarized in Table 7 and are generally in line with the pattern of results reported for informative task 1.With one exception, the correlations between corresponding latent variables (i.e., content variables based on holistic versus analytic scoring, etc.) were high for all four tasks, ranging between .81 and 1. The results were particularly clear for the language factors which correlated close to 1 for all tasks. Likewise, the semi-holistic content and style ratings were strongly associated with the content and structure variables based on analytic scoring (content, structure, 'semantics/pragmatics') whereas correlations between the language factor from one scoring approach and the content factor based on the other scoring approach were nominally lower. An exception was informative task 2 with a relatively moderate correlation of .63 between the content factors. However, this was a peculiarity of the analytic content factor in this task which had even lower correlations with the other factors, most notably language.

In sum, with regard to research question three the results suggest that both scoring systems measure the same latent constructs. Exceptions from this rule were task-specific and not associated with a particular text genre.

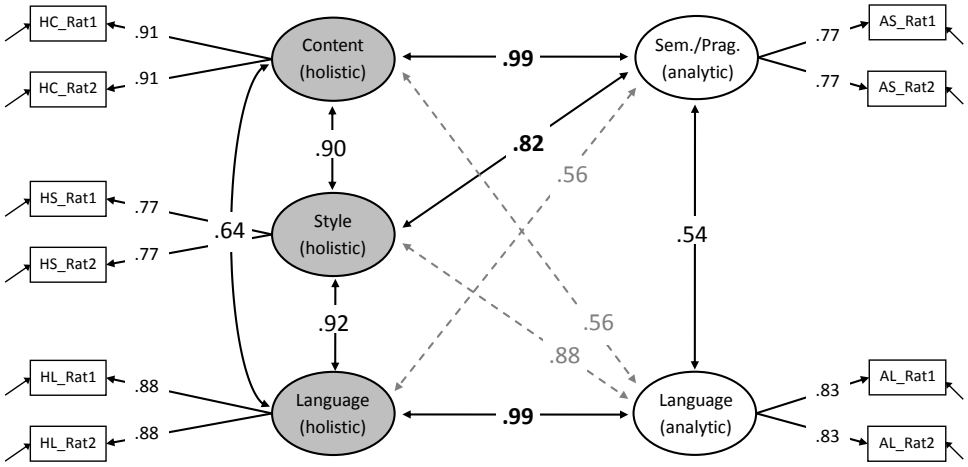

Figure 1. Structural model for informative task 1 with latent variables representing ratings based on two different scoring systems. N = 178. Model fit: χ^2 = 38.1, df = 24, p = .034; CFI = .990, RMSEA = .057, WRMR = 0.731. Sem./Prag. = Semantics/pragmatics, Rat = Rater, H = holistic, A = analytic, C = content, S = style (on the left) or semantics/pragmatics (on the right), L = language

Table 7. Correlations Between Latent Variables Based on the Semi-Holistic Ratings (Left Column) and the Analytic Scoring Procedure (Columns on the Right), Respectively

Semi-holistic scales	Analytic rating procedure			
	Content	Structure	Language	Sem./Prag.
Persuasive task 1				
Content	**.82**	.77	.60	-
Style	.81	**.81**	.73	-
Language correctness	.47	.57	**.96**	-
Persuasive task 2				
Content	-	-	.74	**.88**
Style	-	-	.78	**1.0** [a]
Language correctness	-	-	**1.0** [a]	.74
Informative task 1				
Content	-	-	.56	**.99**
Style	-	-	.88	**.82**
Language correctness	-	-	**.99**	.56
Informative task 2				
Content	**.63**	.69	.69	-
Style	.58	**.91**	.81	-
Language correctness	.37	.70	**1.0** [a]	-

Note. N varies between 163 and 184 depending on the task. Correlations between corresponding variables are printed in bold type. Sem./Prag. = Semantics/pragmatics.
[a] Correlation > 1 when freely estimated.

4. DISCUSSION

In this study, we investigated whether two different scoring systems for the assessment of writing proficiency produce comparable results in terms of their reliability (i.e., inter-rater and intra-rater agreement with trained raters) and validity (i.e., dimensionality of the writing construct as assessed with each scoring system and relationships between the constructs represented by the respective scores).

With regard to our first research question, the results showed acceptable reliability for all scales of the holistic scoring approach (i.e., the global impression scale and the semi-holistic scales). Likewise, most analytic criteria fulfilled basic reliability requirements. However, contrary to reports in the literature (e.g., Weigle, 2002; Knoch, 2009b), the majority of the analytic criteria was less reliable than the holistic and semi-holistic scales. It is important to keep in mind that the term "analytic scoring" in the literature often refers to rating scales similar to the semi-holistic scales analyzed in the present study; in contrast, the dichotomous criteria investigated here could be considered "feature analysis" (Swain, Graves, & Morse, 2010). Despite their more favorable reliability, it remains an open research question whether the holistic and semi-holistic scores are more susceptible to judgemental errors such as the halo effect (Thorndike, 1920) due to the more subjective nature of the judgements (but see Böhme et al., 2009, for an analysis of halo effects in holistic ratings and analytic criteria based on essays from primary school students).

Dimensionality analyses for both scoring systems (see research question two) led to similar results: Whereas content and stylistic or structural aspects were highly associated and difficult to differentiate from an empirical perspective, language correctness was a dimension of writing ability that was distinguishable from the other aspects, although substantially correlated with them. These results are in line with findings for primary education (Böhme et al., 2009) and previous studies in secondary education (Neumann, 2007) for German (L1). Especially for analytic scoring, results were dependent on the task but not the text genre. For two out of four tasks, a model with three factors for content, structure, and language, respectively, was more appropriate whereas for the other two tasks, a simpler model prevailed that did not differentiate between content and structure. This result could be seen as an indication that the analytic criteria showed a non-trivial amount of differential functioning depending on the task and was in line with findings emphasizing the task specificity of writing assesssment results (Bouwer et al., 2015; Kuhlemeier & van den Bergh, 1998; Schoonen, 2005). Note however that in the present study, analytic scoring was conducted text-wise, that is, a given rater rated all analytic criteria for a given text before proceeding to the next student response. Text-wise coding may have resulted in halo effects (i.e., an overestimation of the intercorrelations between the text features).

With regard to research question three, a direct comparison of the latent factors derived from both scoring approaches lead to the conclusion that both evaluation procedures measured essentially the same constructs. This finding coincides with and extends the conclusion of Böhme et al. (2009) who found a strong convergence of holistic and analytic evaluation procedures on the construct level for primary education.

Although both holistic scoring (including the semi-holistic scales) and analytic scoring provided reliable and valid information about the students' writing ability and despite the fact that both scoring approaches converged on the construct level, it could be argued that they are nonetheless not equally suited for all kinds of assessments. Hence, as point-

ed out above, the purpose of the writing assessment should be taken into consideration when deciding on a scoring strategy (Bacha, 2001; Knoch, 2011). An overall judgment like the global impression rating used in this study provides a comprehensive, reliable, and economic quantification of a student's writing ability. For assessment programs such as PISA which report on a system level only, this information might be sufficient. However, a single score that attests an insufficient level of competence could not be used for detailed diagnosis and intervention in the classroom setting. For this purpose, one alternative is the use of semi-holistic variables sensitive to specific aspects of writing such as content or language correctness. Analytic criteria (i.e., feature analyses) promise an even more detailed description of the student's writing proficiency and identification of particular deficits as a basis for effective support. However, our results indicate that this conceptual advantage of analytic scoring might be strongly reduced in practice due to relatively low inter-rater agreement, making individual feedback on the basis of single criteria questionable. A possible solution to reliability issues is the aggregation of several criteria into scores, for instance the calculation of sum scores for content, structure, and language, respectively. In the present study, we found that reliability of such composite scores was on par with the (semi-)holistic scales. However, the theoretical advantage of the analytic approach over the (semi-) holistic scales is in the specificity and heterogeneity of the criteria. Calculating sum scores nullifies this advantage.

Besides reliability, a second limitation for the use of writing scores that seems to be more pronounced for analytic criteria is task specificity. From a diagnostic perspective, the dependency of scores on the specific writing task (or set of tasks) threatens the generalizability of the results and is especially critical for individual feedback. For conclusions on a superordinate level (e.g., average level of writing proficiency in a state or country), one can attenuate the problem by using a large variety of writing tasks in a complex test design and calculating a single score based on all tasks. This strategy, however, does not fit the demand of instructors for detailed feedback on individual strengths and weaknesses. If each student completes only a single task or very few writing tasks from the same text genre and one aims at providing individual feedback, a different approach may be more promising, such as giving genre-specific or even task-specific feedback that sheds light on whether or to what extent the learning aims for a class of tasks or a specific task were met (Bachman, 2002; Bouwer et al., 2015; Olinghouse, Santangelo, & Wilson, 2012).

The improvement of existing scoring procedures to provide a feasible compromise between reliability, cost effectiveness, and detailed feedback remains an important area of academic research. An open research question that should be the focus of future studies on the scoring of student essays pertains to the generalizability of the results presented here to the scoring of longer student texts (e.g., writing exam essays at the end of secondary education). However, in the present study we focused on large-scale assessment contexts in which the use of extensive writing tasks is unusual due to limitations in testing time and the rationale to measure a complex construct such as writing ability with a variety of tasks (Wittmann, 2002). Another limitation of the current study arises from the relatively limited number of writing tasks and the complex booklet design that permitted only two writing tasks per student. Future research should include more tasks and text genres (e.g., narrative writing) to analyze task and genre effects more systematically and extensively (see e.g., Bouwer et al., 2015; Schoonen, 2005, 2012). Finally, while instructors and other experts on didactics frequently demand detailed feedback as a basis for interven-

tions, it is still largely unclear which specific interventions that are tailored to the individual student are most effective in promoting writing proficiency or simply more effective than general interventions such as providing more writing opportunities (but see Graham & Perin, 2007).

To conclude, with regard to our main research questions the results of the present study can be summarized as follows: (i) The analytic and the holistic scoring approach—as implemented here—measure the same constructs. (ii) Both scoring procedures tap two major dimensions of writing (i.e., language correctness on the one hand and content/style on the other hand), which suggests that both provide valid indicators of writing ability. (iii) Both scoring procedures provide reliable estimates on a global level (integrated total score or global impression) and for the major dimensions (composite scores or semi-holistic scales). However, the analytic approach could not play out its conceptual advantage—that is, detailed feedback on specific aspects of writing—since many of the analytic criteria were not reliable enough for individual feedback. With regard to large-scale assessments, the results imply that for the purpose of system monitoring on the state level, the holistic global scale seems most adequate. For diagnostic assessments that aim at improving classroom practices, the semi-holistic scales provide reliable information on major dimensions of writing. Of course, it is also possible to use composite scores based on analytic criteria; however, the rating scales proved to be more efficient and essentially provided the same information.

ACKNOWLEDGEMENTS

During the preparation of this manuscript, Stefan Schipolowski was a fellow of the International Max Planck Research School “The Life Course: Evolutionary and Ontogenetic Dynamics (LIFE)

REFERENCES

Alderson, C. (2006). *Diagnosing foreign language proficiency: the interface between learning and assessment*. London: Continuum.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Available from http://www.jtla.org

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*, 125-141. http://dx.doi.org/10.1177/0265532212452396

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

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System, 29*, 371-383. http://dx.doi.org/10.1016/S0346-251X(01)00025-2

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*, 453-476. http://dx.doi.org/10.1191/0265532202lt240oa

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bae, J., & Bachman, L. F. (2010). An investigation of four writing traits and two tasks across two languages. *Language Testing, 27*, 213-234. http://dx.doi.org/10.1177/0265532209349470

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*, 49-58.

Böhme, K., Bremerich-Vos, A., & Robitzsch, A. (2009). Aspekte der Kodierung von Schreibaufgaben [Aspects of essay scoring]. In D. Granzer, O. Köller, & A. Bremerich-Vos (Eds.), *Bildungsstandards Deutsch und Mathema-*

tik. Leistungsmessung in der Grundschule [National educational standards for German and mathematics. Performance assessment in primary school] (pp. 290-329). Weinheim: Beltz.

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing, 32*, 83-100. http://dx.doi.org/10.1177/0265532214542994

Bryant, B. R., & Bryant, D. P. (2003). Assessing the writing abilities and instructional needs of students. In C. R. Reynolds, & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: intelligence, aptitude and achievement* (pp. 419-437). New York: Guilford Publications.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*, 163-178. http://dx.doi.org/10.1111/j.1745-3984.2000.tb01081.x

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3-32). Urbana, IL: NCTE.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*, 5-43. http://dx.doi.org/10.1016/j.asw.2005.02.001

Eckes, T. (2008). Rater types in writing performance assessments: a classification approach to rater variability. *Language Testing, 25*, 155-185. http://dx.doi.org/10.1177/0265532207086780

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical Methods for Rates and Proportions, Third Edition*. Hoboken, NJ: John Wiley & Sons, Inc.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: theory and practice. *Educational Measurement: Issues and Practice, 28*, 39-53. http://dx.doi.org/10.1111/j.1745-3992.2009.00154.x

Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters scores for speeches. *Journal of Research and Development in Education, 27*, 73-82.

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

Huot, B. (1990a). Reliability, validity, and holistic scoring: what we know and what we need to know. *College Composition and Communication, 41*, 201-213.

Huot, B. (1990b). The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research, 60*, 237-263.

KMK (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring [Strategy of the Standing Conference of Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany for quality development in education].* München: Luchterhand/Wolters Kluwer.

KMK (2014). *Basic structure of the education system in the Federal Republic of Germany*. *Diagram*. Retrieved from http://www.kmk.org/information-in-english/the-education-system-in-the-federal-republic-of-germany.html

Knoch, U. (2009a). Diagnostic assessment of writing: a comparison of two rating scales. *Language Testing, 26*, 275-304. http://dx.doi.org/10.1177/0265532208101008

Knoch, U. (2009b). *Diagnostic writing assessment: the development and validation of a rating scale*. Frankfurt/Main: Peter Lang.

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*, 81-96. http://dx.doi.org/10.1016/j.asw.2011.02.003

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: how does it compare with face-to-face training? *Assessing Writing, 12*, 26-43. http://dx.doi.org/10.1016/j.asw.2007.04.001

Köller, O., Knigge, M., & Tesch, B. (2010). (Eds.). *Sprachliche Kompetenzen im Ländervergleich [National assessment of language competences]*. Münster: Waxmann.

Kuhlemeier, H., & van den Bergh, H. (1998). Relationships between language skills and task effects. *Perceptual and Motor Skills, 86*, 443-463.

Lenhard, W., Baier, H., Hoffmann, J., & Schneider, W. (2007). Automatische Bewertung offener Antworten mittels Latenter Semantischer Analyse [Automated scoring of open responses using latent semantic analysis]. *Diagnostica, 53*, 155-165. http://dx.doi.org/10.1026/0012-1924.53.3.155

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper, & L. Odell (Eds.), *Evaluating writing: describing, measuring, judging* (pp. 33-66). Buffalo: State University of New York.

Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus user's guide. Fifth edition*. Los Angeles, CA: Muthén & Muthén.

Neumann, A. (2007). *Briefe schreiben in Klasse 9 und 11 [Writing letters in Grades 9 and 11].* Münster: Waxmann.

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

OECD (2009). *PISA 2009 assessment framework - key competencies in reading, mathematics and science*. Paris: OECD Publishing.

Olinghouse, N. G., Santangelo, T., & Wilson, J. (2012). Examining the validity of single-occasion, single-genre, holistically scored writing assessments. In Steendam, E. V., Tillema, M., Rijlaarsdam, G., & van den Bergh, H. (Eds.), *Measuring writing: recent insights into theory, methodology and practices* (pp. 55-82). Leiden & Boston: Brill.

Pant, H. A., Gärtner, H., Harych, P., Kuhl, P., & Wendt, W. (2008). Die Evaluation schulischer Bildungserträge auf Länderebene. Das Institut für Schulqualität der Länder Berlin und Brandenburg (ISQ) [Evaluation of educational achievement on the state level. The Institute for School Quality of the Länder Berlin and Brandenburg (ISQ)]. *Zeitschrift für Evaluation, 7*, 309-322.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Eds.). (2013). *The IQB National Assessment Study 2012. Summary.* Available from https://www.iqb.hu-berlin.de/laendervergleich/lv2012/Bericht

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

Rijlaarsdam, G., Van den Bergh, H., Couzijn, M., Janssen, T., Braaksma, M., Tillema, M., Van Steendam, E., Raedts, M. (2012). Writing. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook – Vol. 3: Application to learning and teaching* (pp. 189-227). Washington, DC: APA.

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

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

Shermis, M. D., & Burstein, J. C. (2013). *Handbook on automated essay evaluation: current applications and new directions*. New York, NY: Routledge.

Shohamy, E., Gordon, C. M., Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33. http://dx.doi.org/10.1111/j.1540-4781.1992.tb02574.x

Slater, S. C. & Boulet, J. R. (2001). Predicting holistic ratings of written performance assessments from analytic scoring. *Advances in Health Sciences Education, 6*, 103-119.

Stanat, P. & Christensen, G. (2006). *Where immigrant students succeed. A comparative review of performance and engagement in PISA 2003*. Paris: OECD.

Swain, S. S., Graves, R. L., & Morse, D. T. (2010). Prominent feature analysis: what it means for the classroom. *English Journal, 99*(4), 84-89.

Thorndike, E. L. (1920). A constant error on psychological rating. Journal of Applied Psychology, 4, 25-29.

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

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263-287. http://dx.doi.org/10.1177/026553229801500205

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

Wittmann, W. W. (2002). Brunswik-Symmetrie: Ein Schlüsselkonzept für erfolgreiche psychologische Forschung [Brunswik symmetry: a key concept for successful psychological research]. In M. Myrtek (Ed.), *Die Person im biologischen und sozialen Kontext [The individual in biological and social context]* (pp. 163-186). Göttingen: Hogrefe.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*, 2-13. http://dx.doi.org/10.1111/j.1745-3992.2011.00223.x

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität [Rater agreement and rater reliability]*. Göttingen: Hogrefe.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Doctoral dissertation, University of California, Los Angeles.