frontiers | Frontiers in Education

ORIGINAL RESEARCH
published: 15 June 2022
doi: 10.3389/feduc.2022.919756

Generalizability of Written Expression Curriculum-Based-Measurement in the German Language: What Are the Major Sources of Variability?

Julia Winkes[1]*†‡ *and Pascale Schaller*[2]†‡

[1] *Department of Special Education, University of Fribourg, Fribourg, Switzerland,* [2] *Institute of Primary Education, University of Teacher Education Bern, Bern, Switzerland*

This study aimed to identify the sources of measurement error that contribute to the intraindividual variability of written expression curriculum-based measurement (CBM-W) and assess how many German writing samples of 3 or 5 min duration are necessary to make sufficiently reliable relative and absolute decisions. Students in grade 3 (N = 128) and grade 6 (N = 118) wrote five CBM-W probes of 5 min each within 1 week, which were scored for commonly used metrics (i.e., words written, correct writing sequences). Analyses within the generalizability theory framework showed that between-student differences accounted for 36–60% of the variance. The student × writing prompt interaction was the largest source of variability, particularly among younger students (44%), while writing prompt *per se* and writing time explained no variance. Two to four writing samples of 3 min are sufficient for most scoring methods to achieve relative reliability $>$0.80. CBM-W in German proved inadequate for the grade levels studied for absolute decisions. These findings imply that CBM-W in this form in German-speaking primary grades is suitable as a universal screening tool but not as a tool for progress monitoring of individual students.

Keywords: curriculum-based measurement (CBM), writing, generalizability theory, reliability, variability

OPEN ACCESS

Edited by:
Markus Gebhardt,
University of Regensburg, Germany

Reviewed by:
Sterett Mercer,
The University of British Columbia, Canada
Stefanie Roos,
University of Siegen, Germany

****Correspondence:***
Julia Winkes
julia.winkes@unifr.ch

†ORCID:
Julia Winkes
orcid.org/0000-0002-4383-1926
Pascale Schaller
orcid.org/0000-0002-3600-5386

‡These authors have contributed equally to this work

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *13 April 2022*
Accepted: *23 May 2022*
Published: *15 June 2022*

Citation:
Winkes J and Schaller P (2022) Generalizability of Written Expression Curriculum-Based-Measurement in the German Language: What Are the Major Sources of Variability? Front. Educ. 7:919756. doi: 10.3389/feduc.2022.919756

INTRODUCTION

Although writing is a crucial competence for students' academic and professional success (Traga Philippakos and FitzPatrick, 2018, p. 165), The National Commission on Writing in America's Schools and Colleges (2003) designated it a neglected basic skill. This wake-up call was a response to the National Assessments of Educational Progress, which captured many students who did not reach a proficient writing level. In 2011, for example, 74% of eighth-graders scored at the "basic" or "below basic" levels, and only 3% could be described as advanced writers compared to their grade level requirements (National Center for Education Statistics, 2011). So, in addition to students with a learning disability, there are a significant number of low achieving writers who lack writing proficiency (Graham and Perin, 2007). Until now, writing at the text level has hardly been included in national assessments in German speaking countries, with the exception of the DESI study (DESI-Konsortium, 2006). It showed that at grade 9, about 29% of the students are not able to formulate a letter adequately for the addressee and that the linguistic quality of these students' writing is also extremely low. Thus, although the educational system, curriculum, teaching methods, and orthography to be learned differ between German-speaking and English-speaking countries, it can be surmised that, as in English-speaking countries, weak writing skills are present but probably underdiagnosed in German-speaking countries. The problem is exacerbated by the lack of standardized writing assessments in German, so that writing is usually only systematically evaluated at the spelling level and cannot be reliably assessed at the text level. Struggling writers produce texts that are generally shorter, less interesting, and poorly organized at the sentence and paragraph level (Hooper et al., 2002). The children's texts are marred by inordinate numbers of mechanical, spelling, and grammatical errors (Dockrell et al., 2015). Therefore, the difficulties of these children go far beyond pure spelling problems since the spelling is only a small part of the skills required to produce linguistically correct and content-appropriate texts of good quality. A competency that is an indicator of writing quality at the text level is writing fluency (Kim et al., 2017; Poch et al., 2021). At the same time, writing fluency proves to be sensitive to change since both speed/productivity and accuracy increase with a growing writing routine. Skills that serve as indicators of general performance in an academic area are useful as vital signs for screening students at risk and for progress monitoring (Fuchs, 2004, 2017). For this purpose, short, reliable, and valid learning samples are used in curriculum-based measurements (CBM), which capture critical skills simply and economically (Deno, 1985).

CBM Writing (CBM-W), as an indicator of writing proficiency, uses short writing samples for this aim. The students are given writing prompts, such as pictures or introductory sentences and asked to write for 3 or 5 min. Various scoring methods are available, such as the number of words written (TWW), the number of words spelled correctly (WSC), the number of correct writing sequences (CWS), or the number of correct minus incorrect writing sequences (CIWS). Thus, the collected measures do not focus on content-related text quality (e.g., ideation or genre specificity) but either on writing quantity (TWW), spelling (WSC), or linguistic units whose correct realization requires the integration of individual sub-competencies (writing motor skills, retrieval of linguistic knowledge, semantics and spelling) (CWS and CIWS).

Since the beginning of CBM research, great importance has been attached to ensuring that the methods used reflect the learners' performance in reliable ways – despite their easy handling and the short time for implementation and scoring (Fuchs et al., 1983). Reliable information is key because it builds the foundation for the teachers' important (high stakes and low stakes) data-based decisions (McMaster and Espin, 2007). Parallel forms are needed in their functions as repeated screenings and progress monitoring. These require high parallel test reliability (rank-ordering of students) and stability (consistent within-student performance over time) (Campbell et al., 2013). This central claim contrasts with an observation we made in a previous research project on CBM-W (Winkes and Schaller, 2022). In this study, students in grades 3–6 wrote ten writing samples within a short period of 2 weeks. Parallel test reliability was satisfactory overall, but a closer look at the children's test data revealed considerable intraindividual variability between student test scores. We found this observation remarkable because the CBM samples were collected within a quite short period. In general, meaningful variation in performance within individuals is not fundamentally new for CBM (Christ et al., 2016). It invites a closer look at the issue of "variability" – here specific to CBM-W. Accordingly, the present study aims to understand the sources of this variability in more detail and examine the influence of story starter, rater/class, and length of writing sample on the generalizability of CBM-W in German.

Potential Sources of Variability in Written Expression Curriculum-Based Measurement

Taking the object of learning as a starting point, increased intraindividual variability in writing, compared to other performance areas such as reading, spelling, or mathematics, is not necessarily surprising. On the one hand, variability can be understood as an expression of the complexity of the writing process itself. Text writing is a problem-solving process that requires the integration of different hierarchy-low and hierarchy-high processing skills and thus does not succeed with equal fluency and quality at all times (Alamargot and Chanquoy, 2001; Kent and Wanzek, 2016). On the other hand, writing a text is a creative language-productive task, which leads to a special starting point. In other areas of CBM, the number of given items (e.g., arithmetic problems, words to be read) that can be correctly solved in a defined time is usually recorded. In writing, on the other hand, the items to be assessed are produced by the child himself.

Two children with the same writing skills will arrive at two very different final products based on the same story starter. The same is true when testing a child repeatedly. Even using the same story starter and under comparable contextual conditions, a child is unlikely to use the same words and phrases to write a story on two different occasions. As Ritchey et al. (2016) point out, writing opens up opportunities for students to actively avoid difficult words or choose simpler words and sentence structures, which influences the difficulty of different texts.

A certain variability is, therefore, to be expected, which is inherent to the writing process itself and which is caused by the open nature of the task. For this reason, it is particularly important in writing to design the conditions for progress-monitoring measurements so that as many external sources of measurement error as possible can be reduced and that as much of the remaining variance as possible can be attributed to the subject itself. Potential sources of measurement error could include, for example, the different story starters, the length of the writing time, or the rater. In the following, we discuss the state of knowledge regarding the importance of these factors concerning CBM-W.

The Task or Writing Prompt

So far, the role of writing prompts has been surprisingly little investigated in CBM-W. Existing studies on this topic focus on what kinds of writing prompts are appropriate at which grade level. For example, various word- and sentence-level task formats have been suggested for beginning writers, requiring text production in response to a picture or story starter with descriptive or narrative content (Ritchey et al., 2016). In the higher grades, the question arises in particular whether expository or narrative prompts better represent students' academic writing abilities, as they are potentially more in line with typical school writing tasks [for two recent meta-analyses related to the validity of different writing genres, see Romig et al. (2017, 2020)]. Within the different genres (e.g., expository vs. narrative), it is assumed that different tasks are comparable and that the writing prompts used are equivalent, without this assumption having been sufficiently tested empirically to date (Keller-Margulis et al., 2016a). In contrast, for other forms of CBM (e.g., reading fluency; mathematics), great importance is attached to the development of parallel test versions. As Christ et al. (2016) describe, the variability of student performance across forms in CBM research has led to the standardization not only of the procedures for administration and scoring but also of the materials used. This development does not seem to have established itself specifically in writing. While collections of tasks are available at CBM-W[1, 2], it is also possible for practitioners to invent story starters themselves, as long as they are age-appropriate and do not evoke a one-word response (Hosp et al., 2016). However, McMaster and Espin (2007) point out that students' background knowledge and interest in different writing prompts may vary greatly, affecting the quality and quantity of their writing. Existing studies of writing prompt comparability use alternate-form reliability to examine how closely different writing samples correlate with each other [see for grades 1–5 the studies of Gansle et al. (2002), Weissenburger and Espin (2005), Gansle et al. (2006), Campbell et al. (2013), and Allen et al. (2019)]. They usually set a Pearson's correlation coefficient of $r \geq 0.70$ for sufficient reliability in CBM-W (Allen et al., 2019, p. 10). The various scores usually reach this threshold. However, Allen et al. (2019) found large differences between the correlation coefficients. For example, for grade level 3, the CIWS coefficients vary between 0.31 and 0.92, and for TWW, between 0.50 and 0.91. McMaster and Espin (2007, p. 69) point out that the standards for reliability coefficients should possibly be set domain-dependently. For CBM of oral reading fluency, reliability coefficients of $r > 0.85$ are usually reported. Such high coefficients are not expected for CBM-W, which is probably related to the test setting: A text as a continuation of a story starter can take an infinite number of possible forms, which is not the case for reading fluency. Moreover, the procedure established in CBM-W for eliciting parallel test reliability, namely calculating the correlations between several CBM tests administered simultaneously, only verifies part of the necessary conditions for parallel tests. These should also have equal means and variances (Christ and Hintze, 2007). This assumption has not yet been controlled for CBM-W.

The Writing Time

The main characteristic of progress monitoring and CBM procedures is that they are highly time-efficient in implementation and evaluation (Deno, 2003). This is the only way to ensure that regular use is possible in everyday school life, especially if used in parallel in several performance areas (e.g., reading, spelling, writing, mathematics). Thus, the duration of CBM-W should be as short as possible but as long as necessary to ensure a sufficiently reliable capture of the feature to be examined. Most studies on CBM-W refer to 3-min writing samples preceded by a planning period of 1 min, and this procedure is also the standard in practice (Hosp and Kaldenberg, 2020). However, the effects of increased writing time (e.g., 5, 7, or 10 min) on the reliability of measures in CBM-W have been studied on several occasions. Younger students showed only slight differences in the reliability of shorter and longer writing samples (Espin et al., 2000). For older students, increasing the writing time to 5–7 min was necessary to achieve reliability > 0.70 (Weissenburger and Espin, 2005; Campbell et al., 2013), which was also true for the English language learners (ELL; Espin et al., 2008). It is still unclear up to which grade level a writing time of 3 min is sufficient and from when the writing time should be increased. Of course, the choice of writing duration also depends on the purpose. Espin et al. (2008) recommend a 7-min writing sample for older students due to increased reliability if CBM-W is used as a screening only one to three times per school year. For use at shorter and more regular intervals (e.g., once per week), they recommend a more economical 5-min writing sample.

The Rater

Since CBM-W evaluates texts using different scores, the question arises as to what role the rater's influence plays in the results. Campbell et al. (2013) report very high interrater reliabilities: they indicate average interscorer agreement from 80% (CIWS) to 99% (TWW). The differences between scores that report text volume (TWW) and scores that address writing accuracy can plausibly be explained because, in TWW, only the words are counted, whereas CWS or CIWS assess the correctness of writing sequences. Different ratings of the same writing sequence are sometimes related to the fact that different raters assume different target structures announced by the child. The very high interrater reliabilities also for CWS and CIWS [see, Weissenburger and Espin (2005), Gansle et al. (2006), Campbell et al. (2013), and Keller-Margulis et al. (2016b)] are probably due to intensive training of raters, which cannot be assumed in the practical application of CBM-W.

Generalizability Theory

Studies on the psychometric properties of CBM-W have so far almost exclusively used the framework of Classical Test Theory (CTT) by investigating parameters such as parallel test reliability, interrater reliability, or criterion validity. Especially in the context of progress monitoring, where an idiographic reference norm is usually used, Generalizability Theory (G-theory) provides an alternative. It has three advantages: First, it can investigate different sources of measurement error simultaneously. It uses repeated measures ANOVA to estimate the variance components for each source of variation (referred to as facets in G-theory terminology) in the observed values and the interactions among these facets. Thus, G-theory provides a good overview of the main contributors to measurement error, which, unlike in CTT, are analyzed in the same model. This information

[1] www.aimsweb.com

[2] www.interventioncentral.org

can subsequently be used to effectively optimize assessment procedures (Hintze et al., 2000).

The second advantage of G-theory concerns the reliability coefficients reported. In CTT, the calculation of parallel test reliability examines whether a child moves in the same rank relative to the other children in the group on repeated performance measures within the subject group, such as the class. If, for example, the weakest child in the class always achieves the lowest measurement result in the class over five measurement points, then classical test theory evaluates this as an indication of high parallel test reliability, although the child's competence values may vary greatly between these five measurement points (see Keller-Margulis et al., 2016a). In G-theory, there is a corresponding coefficient of generalizability (G-coefficient) to this classical reliability coefficient, which is thus informative for relative decisions related to the ranking of subjects (Cardinet, 1998).

In addition, the dependability-coefficient (D-Coefficient) is another parameter that focuses on the performance level, independent of the ranking. It can be used to make absolute, criterion-referenced decisions. D-coefficients are more conservative than G-coefficients for this reason. They are particularly suitable for use in progress monitoring, as Fan and Hansmann (2015) argue: "... research has acknowledged that having high-rank order reliability at a group design level (like the generalizability coefficient in G theory) cannot guarantee the comparability of CBM-R scores used at the individual student level" (S. 207). The minimum thresholds of G- and D-coefficients depend on the application situation. For low-stakes decisions, a reliability of 0.80 is considered sufficient and feasible in practice. However, for high-stakes decisions it is usually argued referring to Nunnally (1967) that coefficients below 0.90 are unacceptable (Graham et al., 2016; Keller-Margulis et al., 2016a; Kim et al., 2017; Wilson et al., 2019). The third advantage of G-theory is that G-coefficients and D-coefficients can not only be generated for the actual conditions of investigation but it can be estimated with the help of so-called decision studies (D-studies) how these coefficients vary under other conditions. This allows identifying the minimum requirements to obtain sufficiently high measurement reliability. For example, how many writing samples of what duration are necessary to achieve reliability above 0.80 for relative or absolute decisions can be checked.

Use of Generalizability Theory in Written Expression Curriculum-Based Measurement

The advantages of G-theory over CTT lead to popularity in writing assessments. A recent review of the content of the journal "Assessing Writing" from 2000 to 2018 (Zheng and Yu, 2019) indicates that G-theory was the most frequently used method during this period. However, existing studies mainly examined college students or adult L2 learners. Which factors influence the reliability of writing scores in children has not yet been much addressed (Kim et al., 2017). Specifically, for CBM-W, generalizability theory has been used only twice: In the study of Keller-Margulis et al. (2016a), 2nd–5th grade students wrote three 7-min writing samples at three time points each year. After each minute, subjects changed the color of their pen while writing so that the impact of writing time on the reliability of the measures could be assessed (from 1 to 7 min). Other facets included students (between student differences), story starter, benchmark (time within a year), and interactions among these factors. Nearly half of the variance in CBM-W proved to be the non-systematic error. Reliability above 0.80 – as the threshold for low-stakes decisions – was achieved with the relative reliability coefficient at most grade levels by three 3-min writing samples, the D-coefficient for absolute decisions reached the threshold of 0.80 with two 5-min or three 4-min tests. For contexts with high stakes decisions, depending on grade level and scoring method, three 5- to 7-min writing samples were needed for sufficient relative reliability above 0.90, and three 7-min writing samples were necessary for sufficient absolute reliability. Thus, the typical CBM-W implementation convention of using a single writing sample of 3 min as a screening instrument proves inadequate. The use of multiple longer writing samples, on the other hand, severely limits the feasibility of CBM-W in its function as a screening, making widespread implementation unrealistic for many schools. Therefore, the authors are skeptical about whether CBM-W is the best way to identify at-risk students in writing.

In the second study, which used G-theory, Kim et al. (2017) examined the influence of rater ($N = 2$) and task ($N = 3$) on the reliability of writing tasks in expository and narrative genres for 3rd and 4th-grade students. The writing time here was 15 min per text, so the task does not correspond to conventional implementation conditions for CBM-W, but the texts were analyzed using the scoring methods for CBM-W, among others. For the evaluation *via* TWW and CWS, it was found that most of the variance was explained by the person (57–69%) and another large proportion by the interaction between person and task (31–41%). Variability was minimal when explained by rater, person × rater, or the non-systematic error. Subsequent D-studies indicated that for both absolute and relative decisions, two to four tasks and a single rater were necessary to reach the criterion of 0.80 and five to six tasks and one rater were necessary for the criterion of 0.90 reliability.

The Present Study

The present study explores the major sources of variability of CBM-W in German in grades 3 and 6. CBM-W has only been investigated in two studies with divergent results using G-theory. Language structural differences also prevent the unreflected transfer of evaluation measures from one language to another: While the English orthography has a deep phoneme-grapheme correspondence, German has a more complex morphemic structure than English, which affects word length. German also has more complex rules for capitalization and punctuation (commas). Due to these linguistic differences, it is important to go beyond existing English-language studies to determine the optimal conditions for CBM-W in German.

The two central questions for the practical application of CBM-W, which we will address in the planned paper, are:

(1) Which factors contribute to intraindividual variability in CBM-W, and to what extent?
(2) Under which minimum measurement conditions does CBM-W achieve sufficient reliability for relative and absolute decisions?

The following hypotheses precede the data analyses:

(1) In grade 3, prompts play a larger role, meaning that the facet story starter explains more variance than in grade 6. These differences are likely related to the fact that grade 3 children have less extensive vocabularies for certain topics and less world knowledge than grade 6 children. This, in turn, results in the younger children producing less text volume as they spend more time finding words and generating ideas. Thus, vocabulary size and vocabulary quality are likely to have less impact on the test score achieved as children get older.
(2) Increasing writing time from 3 to 5 min positively affects measurement reliability at both grade levels, as reflected in higher G- and D-coefficients. In grade 6, this effect is even more positive since existing studies indicate that in lower grades, shorter writing samples are sufficient for reliable values, whereas, in higher grades, a longer writing time is appropriate to achieve adequate values.
(3) Based on the observation that many children's achieved scores vary between measurement time points, it can be assumed that the D-coefficients differ significantly from the G-coefficients.

MATERIALS AND METHODS

Participants

Written expression curriculum-based measurement was conducted with a sample of third (N = 128) and sixth (N = 118) grade German-speaking students. Nine third grade classes and seven sixth grades classes from nine different schools participated in the study. The participating schools were spread over the German-speaking part of the canton of Fribourg (CH). Schools from both rural and urban areas participated in the study. The sample consisted of 71 girls (55.5%) and 57 boys (44.5%) in grade level 3 and 57 girls (48.3%) and 61 boys (51.7%) in grade level 6. A total of 119 students (48.4%) reported being multilingual, with 163 participants (66.2%) describing German as their first language. On average, the students were 8.8 years (SD 4.4) old in grade 3 and 11.8 years (SD 5.0) in grade 6. The active consent of the Education Directorate of the Canton of Fribourg, the school administrators, the class teacher, the parents and the child was a prerequisite for participation in the study.

Instrument

The instrument consists of five writing samples. The following story starters were used: "Last week I was allowed to take my pet to school when...", "I never believed in magic until Luke at school today...", "While walking on the beach, I discovered a stranded message in a bottle.", "Finally it worked, I invented the machine that...", "My feet are lifting off the ground. I'm flying!". These five writing prompts were used in the same order for the third and sixth grades.

Procedure

The data collection was part of a larger study of writing fluency and its subcomponents.

Administration

The CBM-W samples were collected using a standardized implementation guide by teachers in the participating classes according to the usual standard for conducting CBM-W (Hosp et al., 2016). Students were given a sheet with the pre-printed story starter and lines to write on. They were told they had 1 min to think and then 5 min to write a story. After writing for 3 min, students were asked to mark with a cross the point to which they had written up to that point. The test administrator checked for accurate adherence to the time constraints. The students wrote the five writing samples within one school week.

The evaluations of the tests were done by trained students of special education. The training of the raters included an introduction to the scoring methods and the joint evaluation of several sample texts. There was the possibility to ask questions *via* an online forum during the data evaluation, which was actively used. No systematic checks were made to see if raters agreed with each other. In many other studies on writing assessment, training continues until high interrater reliability is ensured. Error variance attributable to the facet rater can thus be significantly reduced. The procedure chosen here realistically corresponds to the conditions under which CBM-W is implemented in school practice. The influence of the rater is presumably higher in school than in controlled studies, where many hours of rater training time are invested (Allen et al., 2019). Kim et al. (2017) also discuss that in a study examining factors influencing the reliability of a measurement method, it is preferable not to ensure a predetermined level of agreement between raters because the goal is to survey the influence of the facet rater under training conditions that are realistic in practice.

Scoring

This article focuses on four scoring methods: TWW, CWS, CIWS, and %CWS. These scoring methods include production-dependent measures (TWW, CWS), production-independent accuracy measures (%CWS) and accurate-production indices (CIWS) (Malecki and Jewell, 2003; Jewell and Malecki, 2005):

- Total Words Written (TWW): The number of written words separated by another by a space is counted. The words do not have to be spelled correctly (Espin et al., 2000).
- CWS: Fuchs and Fuchs (2007, 12) define CWS as follows: "A correct word sequence is one that contains any two adjacent, correctly spelled words that are acceptable within the context of the same to a native (English) speaker. The term 'acceptable' means that a native speaker would judge the word sequences as syntactically and semantically correct." Thus, the orthographic, semantic, and grammatical fit of what is written is assessed when

evaluating writing sequences. Correct writing sequences are marked with a carat between the two words. The evaluation of correct punctuation in English includes only the correct capital letter at the beginning of the sentence and the correct end mark at the end of the sentence. In German, we also evaluate the presence of necessary commas but not literal speech marks. In addition, it should be noted that in German, all nouns are capitalized, so capitalization is more complex than in English.

- Correct Minus Incorrect Writing Sequences (CIWS): Analogous to the correct writing sequences, incorrect writing sequences can also be evaluated. Between two words or a word and a punctuation mark, an incorrect sequence is then marked using an inverted carat if at least one of the two is incorrect in terms of orthography, semantics, or syntax. Missing elements (words or punctuation marks) in the present study were marked by two consecutive incorrect sequences. Subtracting the incorrect sequences from the correct ones yields an accurate production index, which incorporates writing fluency and accuracy (Jewell and Malecki, 2005).
- Percentage of Correct Writing Sequences (%CWS): This method – calculated as the percentage of correct sequences from the sum of correct and incorrect sequences – is independent of the amount of text written and is therefore considered a measure of accuracy (McMaster and Espin, 2007).

It should be noted that not every scoring method has proven to be equally reliable and valid at every grade level. While TWW is more suitable for younger students at the beginning of writing acquisition, CWS and CIWS are recommended for use around the third-grade level, but certainly for older students (McMaster and Espin, 2007; Saddler and Asaro-Saddler, 2013; McMaster et al., 2017; Romig et al., 2017; Payan et al., 2019).

Data Analysis

The statistical tests include analyses within the framework of G-theory (G-studies, D-studies). All calculations were performed separately for the third and the sixth grade for TWW, CWS, CIWS, and %CWS. The analyses were performed with the software G-String VI (Bloch and Norman, 2021)[3], which is a graphical user interface for the operation of urGENOVA (Brennan, 2001). In the generalizability studies (G-studies), variance components were estimated for main and interaction effects of the facets student (facet of differentiation; between student differences), rater (differences across raters), and story starter (differences in performance across writing prompts). The resulting two-facet design is not fully crossed because the facet student is nested in raters.

Furthermore, it should be noted that the texts were assigned to the raters by class. This methodological aspect will be addressed in more detail in the discussion, but it is already mentioned here to better understand the data. The facet rater thus also includes the differences between different classes, which is why this facet is labeled rater/class in the results tables.

The G-studies are calculated separately for the scoring methods TWW, CWS, and CIWS for 3 and 5 min of writing. Studies that also integrate duration of assessment as a facet must always collect student performance per minute (e.g., words read correctly per minute, math problems solved correctly per minute), since otherwise, the variance explained is simply a sign of more items solved in more time [see, e.g., Christ et al. (2005) and Keller-Margulis et al. (2016a)]. However, in the current study, student performance was not recorded after every minute but only after 3 vs. 5 min. The only scoring method for which time (differences in writing performance due to writing time) can be integrated as a facet in the G-study is the production-independent scoring procedure %CWS. This results in a three-facet design with the corresponding interactions.

In G-theory, negative variance components may occur. If these are small, they are usually set to zero (Stumpp and Großmann, 2009; Bloch and Norman, 2012; Briesch et al., 2014). In the present study, negative variance components are replaced by zero following this suggestion but are marked in the tables (*). To address research question 2, decision studies (D-studies) were subsequently conducted in G-string. These indicate how generalizability and dependability coefficients change when the measurement conditions vary (Briesch et al., 2014). Reported are both types of coefficients for one to five writing samples with 3- or 5-min writing time.

RESULTS

The descriptive results for all scoring methods and both grade levels are shown in **Table 1**. There is an increase in mean performance between the 3rd and 6th-grade levels for all scoring methods and through the increase in writing time.

Results of the G-Studies

The G-studies addressed the question of which factors contribute to the variability of the evaluated scoring methods and to what extent. **Table 2** documents the variance components for TWW, CWS, and CIWS in grades 3 and 6 for 5-min writing samples. The corresponding results for 3-min writing samples are similar to those presented here. They can be found in **Supplementary Table 1**. Obviously, the facet student explains the most variance for all scoring methods in the third and sixth grades. For 5-min writing samples, between 45% (CWS grade 3) and 64% (CIWS grade 6) turn out to be between-student differences. The rater/class facet also explains a significant portion of the variance, between 7 and 24%. The influence of story starter and the interaction story starter × rater is extremely small in both grade levels and across all scoring methods, with a maximum of 3% variance explanation. Residual variance (i.e., non-systematic error) amounts to between 20 and 43%, whereby CIWS in grade 3 stands out due to a high proportion of error variance.

In the G-study for %CWS, time was included as a facet. **Table 3** shows that also, in this case, the facet of differentiation (student) explains a considerable proportion of the variance: about 35% for 3rd grade and 60% for 6th grade. Duration of Assessment (time) does not explain any variance at either grade level (0.02% each),

[3] https://github.com/G-String-Legacy/G_String/releases/tag/1.0.1/gstring_25.jar

TABLE 1 | Descriptive statistics (M and SD).

		Grade 3				Grade 6			
		3 min		5 min		3 min		5 min	
		M	**SD**	***M***	**SD**	***M***	**SD**	***M***	**SD**
Probe 1	TWW	15.79	7.59	27.87	11.90	31.17	10.49	52.71	16.60
	CWS	7.02	4.81	12.60	7.61	24.29	10.68	40.44	17.61
	CIWS	−4.32	7.29	−7.15	11.15	12.09	14.66	19.34	23.84
	% CWS	39.29	18.53	39.00	15.86	65.70	17.91	64.82	16.99
Probe 2	TWW	18.96	8.02	31.52	13.22	32.29	11.74	53.15	18.86
	CWS	8.98	5.85	14.35	8.40	25.48	12.80	41.97	20.87
	CIWS	−4.31	8.98	−8.31	13.68	12.83	16.27	21.39	25.69
	% CWS	40.65	19.73	40.10	18.02	65.35	18.26	64.98	16.90
Probe 3	TWW	19.77	8.80	33.73	13.41	34.43	11.50	56.97	18.33
	CWS	9.56	5.99	15.54	9.12	27.30	12.47	44.85	20.67
	CIWS	−3.91	9.33	−8.35	14.95	14.37	17.11	23.00	28.00
	% CWS	42.61	19.98	40.34	17.26	67.14	18.36	65.96	17.65
Probe 4	TWW	19.30	9.02	33.60	14.16	34.28	13.90	57.96	20.78
	CWS	8.54	6.06	15.02	9.75	26.42	13.59	44.64	21.03
	CIWS	−5.70	9.30	−9.45	14.98	12.95	17.78	21.63	27.25
	% CWS	36.71	19.63	37.88	18.31	65.45	20.69	64.80	18.80
Probe 5	TWW	20.22	9.51	35.05	15.12	35.26	13.27	59.80	20.34
	CWS	10.10	6.99	16.59	10.85	27.48	13.68	46.89	22.19
	CIWS	−3.79	9.05	−8.54	14.60	13.46	17.72	22.86	29.60
	% CWS	41.50	19.57	39.60	17.17	64.80	19.24	64.51	18.93

TABLE 2 | Results of the G-studies for 5-min writing samples for TWW, CWS, and CIWS.

	Grade 3		Grade 6	
Facet	s^2	% s^2	s^2	% s^2
Results for TWW				
Rater/Class	48.33	24.78	46.36	12.46
Student (nested in Rater/Class)	96.35	49.41	213.38	57.36
Story starter	6.72	3.44	9.02	2.42
Rater/Class × Story starter	3.38	1.73	6.70	1.80
Residual	40.31	20.67	97.05	26.08
Total	195.09	100.03	372.51	100.12
Results for CWS				
Rater/Class	19.37	22.26	79.01	18.40
Student (nested in Rater/Class)	39.99	45.96	253.22	58.97
Story starter	2.08	2.39	5.84	1.36
Rater/Class × Story starter	0.49	0.56	1.57	0.37
Residual	26.04	29.93	89.74	20.90
Total	87.97	101.10	429.38	100.00
Results for CIWS				
Rater/Class	13.61	7.08	93.97	12.87
Student (nested in Rater/Class)	92.94	48.32	472.02	64.63
Story starter	−0.39	0.00	1	0.14
Rater/Class × Story starter	2.16	1.12	−1.48	0*
Residual	83.64	43.48	163.36	22.37
Total	192.35	100.00	730.35	100.00

**Negative variance components were set to zero. The sum may differ from 100 due to rounding.*

nor does Story Starter. Particularly informative for %CWS is the Student × Story starter interaction, which contributes most to variance explanation for 3rd grade (44%) and still accounts for 25% for 6th grade.

Results of the D-Studies

When addressing the question "with how many writing samples of which duration and with which scoring procedures does CBM-W achieve sufficient reliability for relative and absolute decisions?" we arrived at different answers for the two investigated grade levels: The results indicate that for the 6th-grade level, more complex scoring measures are indicated for relative decisions, but for the 3rd-grade level already the production measure TWW, measured by two writing samples of 5 min, is sufficient to exceed the threshold for low-stakes decisions of 0.80 (**Table 4**). Also, for CWS, three 5-min writing samples for the 3rd grade reach the value of 0.82, while CIWS and %CWS turn out to be inappropriate for this grade level. The situation is different at the 6th-grade level: for %CWS two 5-min writing samples reach 0.81, for CIWS and CWS already, two 3-min writing samples also reach 0.81, and for TWW, two 5-min writing samples are indicated. If one sets a stricter threshold of 0.90 for high-stakes decisions, it can be reached for students in grade level 3 only from four 5-min writing samples for TWW. In grade 6, the most time-efficient approach for achieving a relative reliability coefficient > 0.90 would be to collect four 3-min samples using CWS or CIWS. Thus, while for relative decisions, procedures can be identified that are sufficiently reliable for making pedagogical decisions, this is not true for absolute decisions: only in one case is a benchmark of 0.80 reached for low-stakes decisions, and that is at the 6th-grade level for four 3- or 5-min writing samples.

DISCUSSION

Procedures for universal screenings and progress monitoring pursue the goal of reliably and validly recording and documenting the individual learning developments of students over time economically. For this purpose, they require parallel tests that show high stability and consistent within-student performance over time. Observations from a previous study on CBM-W (Winkes and Schaller, 2022) revealed, in contrast to this requirement, significant intraindividual variability in the writing performance of German-speaking primary school children over a short-term data collection period. In the present study, we chose generalizability theory as the methodological framework both to address the question of the big sources of variability for CBM-W and to investigate the effects of this variability on the reliability of CBM-W in terms of relative (rank order) and absolute (criterion-referenced) decisions under different measurement conditions.

So, what are the major sources of variability in CBM-W? On the positive side, a substantial portion of variance can be attributed to students (between student differences), ranging from 36 to 65%, depending on grade level and scoring measures. In grade three, student variance explanation is lower than in grade six, where children explain about 60% of the variance for all scoring methods. For the G-studies without the time facet, the second-largest source of variance is unsystematic error variance (20–43%), followed by rater/class (7–25%). For the production-independent scoring method %CWS, assessment duration could be integrated as an additional facet in the G-study. Here, student-story starter-interaction emerges as the main source of variability in grade 3 (44%), ahead of between-student differences (35%). For sixth-graders, the variance explained by student × story starter was much lower, but still 25%. It is also revealing which factors do not turn out to be a big source of variability, which is the case for story starters, for example. Thus, the very small differences between grade 3 and grade 6 are not significant, and hypothesis 1 (story starter has a more important role in grade 3 than in grade 6) could not be confirmed.

Hypothesis 2 assumed that increasing the writing time would positively affect the G- and D- coefficients. This hypothesis is supported, but the differences in the reliability coefficients between 3 and 5 min writing times are small in many cases. As predicted in hypothesis 3, the D-coefficients, on the other hand, deviates significantly from the G-coefficients. While between two and four writing samples are sufficient for relative decisions to exceed the threshold of 0.80, it is not reached by the D-coefficients for absolute decisions with one single exception (%CWS in 6th grade with four texts).

Which Sources of Variability Can Be Optimized for Written Expression Curriculum-Based Measurement?

Compared to other performance domains, assessments in the area of writing generally suggest an increased intraindividual variability. This is probably due in part to the complex cognitive demands of the writing process and in part to the open-ended tasks used in writing assessments (Kent and Wanzek, 2016; Ritchey et al., 2016). In the present study, approximately 60% of the variance was explained by students for all scoring methods for sixth-grade children and somewhat less for third graders. Other studies that examined children's writing performance using generalizability theory, using conventional evaluation methods (e.g., holistic or analytic teacher ratings), consistently found lower variance explained by the facet "person" [e.g., 10% in the study of Bouwer et al. (2015); 38–46% in the study of Graham et al. (2016) and 23–48% in the study of Schoonen (2012)]. Thus, in this respect, CBM-W is not inferior to other forms of writing assessments, also indicated by Kim et al. (2017).

The role of the writing prompt has been investigated for CBM-W primarily in the context of studies of parallel test reliability. However, these studies are less informative when an idiographic frame of comparison is applied, as is the case for progress monitoring (Christ and Hintze, 2007; Christ et al., 2016). That is why G-theory can make a relevant contribution here as an alternative to CTT. The analyses of variance within the G-studies presented indicate that story starters as a facet hardly explain variance. This finding is congruent with existing studies of writing that used G-theory (Schoonen, 2012; Keller-Margulis et al., 2016a; Wilson et al., 2019). This result may be considered

TABLE 3 | Results of the G-study for %CWS with time as a facet.

	Grade 3		Grade 6	
Facet	**s^2**	**% s^2**	**s^2**	**% s^2**
Rater/Class	26.46	7.62	23.17	6.77
Student (nested in Rater/Class)	124.34	35.83	207.90	60.78
Story starter	0.81	0.23	−0.57	0*
Time	0.08	0.02	0.22	0.02
Rater/Class × Story starter	0.90	0.25	2.10	0.61
Rater/Class × Time	0.21	0.06	−0.08	0*
Student × Story starter	153.02	44.09	86.36	25.25
Student × Time	−1.36	0*	0.49	0.14
Story starter × Time	0.61	0.17	−0.16	0*
Rater/Class × Story starter × Time	0.08	0.02	0.032	0.09
Residual	40.50	11.67	22.14	6.47
Total	347.01	99.96	342.70	100.13

**Negative variance components were set to zero. The sum may differ from 100 due to rounding.*

positive in terms of the practical utility of CBM-W in that as many different story starters as desired can be used by teachers. The story starters do not differ systematically in terms of difficulty.

However, of great practical importance for using CBM-W is the interaction between student and story starter which proved to be a large source of variability when estimating the variance components for the scoring method %CWS. It explained 44% of the variance for the younger children (grade 3) and still 25% for the older children (grade 6). This result is in line with other studies on writing assessment, in which this effect also explained a very significant part of the variance (Schoonen, 2012; Bouwer et al., 2015; Graham et al., 2016). The question arises whether this effect in the mentioned studies is due to the combination of tasks of different genres or whether it also exists within one genre. Bouwer et al. (2015) used 12 texts (3 texts in each of four different genres), which were written at three different data collection points. They were able to show that generalizability of children's writing performance between different genres is not warranted (see also Graham et al., 2016). Writing assessments must therefore either include multiple texts of different genres or the interpretation of their results must be narrowed specifically to the genre used. However, the person × task interaction effect persists even within the same genre, as demonstrated by both Bouwer et al. (2015) and Kim et al. (2017). Specific to CBM-W, results to date have been inconsistent. While Kim et al. (2017) documented a large student × task interaction (both within the narrative genre and within expository genre), one did not occur in Keller-Margulis et al. (2016a). Our results support the assumption that it is not the individual story starters *per se* that contribute to variability but rather that children respond differently to tasks. As a possible explanation, it has been suggested that children's background knowledge and experiences differ concerning different writing tasks (Schoonen, 2005; Kim et al., 2017). Since CBM story starters are usually designed to accommodate the child's background experience (Hosp et al., 2016), this reasoning is not completely convincing. The story starters are very open in their formulations and allow the students to make associations in different directions, which is why the world and background knowledge in a specific area should hardly carry any weight, especially since the content of the story is not the subject of the evaluation, but purely formal linguistic aspects are assessed. Therefore, supplementary explanations for the marked interaction effect between a person and a story starter should be considered. We suspect that, especially for younger children, the specific conditions of the writing assessments might have a significant influence, such as the time of day (morning, afternoon), whether the texts were written before or after recess, and which subjects were taught before, and so on. Furthermore, in the writing domain, motivational processes are considered to be of great importance. It is expected that children's personal and situational interests may vary with different writing stimuli and on different occasions (Troia et al., 2012). The influence of external conditions (e.g., time of day) could be included as an additional facet in future studies to verify this hypothesis.

For methodological reasons, the duration of the writing sample could only be integrated into the G-studies for %CWS. It did not explain any variance here, which is also consistent with the results of Keller-Margulis et al. (2016a), who investigated the influence of this facet on the generalizability of CBM-W more systematically. Thus, in the grade levels studied here, there is no evidence that intraindividual variability in the context of CBM-W is caused by the shortness of the writing sample and could be substantially reduced by longer writing samples. As described above, however, the duration of assessment could play a role in older students' writing (Weissenburger and Espin, 2005; Espin et al., 2008; Campbell et al., 2013).

Discussing the role of the rater is difficult for the current study because the raters were assigned by class and thus confounded with class (see below). Both together turn out to be variance components with a significant influence, explaining up to 25% of the variance. Whether differences between raters or between the performance of different classes in different schools manifest themselves here cannot be decided based on the present results and should thus be addressed in further research. However, we cannot exclude – also due to the somewhat more complex scoring

TABLE 4 | Results of the D-studies for TWW, CWS, CIWS, and %CWS.

		Reliability coefficients for TWW			
		Relative decisions (G-coefficient)		**Absolute decisions (D-coefficient)**	
Grade	***n* probes**	**3 min**	**5 min**	**3 min**	**5 min**
3	1	0.62	0.71	0.46	0.49
	2	0.76	0.82	0.55	0.57
	3	0.83	0.88	0.59	0.60
	4	0.86	0.91	0.62	0.61
	5	0.89	0.92	0.63	0.62
6	1	0.63	0.69	0.50	0.57
	2	0.77	0.81	0.61	0.67
	3	0.84	0.87	0.66	0.72
	4	0.87	0.90	0.68	0.74
	5	0.89	0.92	0.70	0.76
		Reliability coefficients for CWS			
		Relative decisions (G-coefficient)		**Absolute decisions (D-coefficient)**	
Grade	***n* probes**	**3 min**	**5 min**	**3 min**	**5 min**
3	1	0.52	0.61	0.40	0.45
	2	0.68	0.75	0.51	0.54
	3	0.77	0.82	0.56	0.58
	4	0.81	0.86	0.58	0.60
	5	0.84	0.88	0.60	0.61
6	1	0.68	0.74	0.53	0.59
	2	0.81	0.85	0.62	0.66
	3	0.87	0.89	0.65	0.69
	4	0.90	0.92	0.67	0.71
	5	0.92	0.93	0.68	0.72
		Reliability coefficients for CIWS			
		Relative decisions (G-coefficient)		**Absolute decisions (D-coefficient)**	
Grade	***n* probes**	**3 min**	**5 min**	**3 min**	**5 min**
3	1	0.42	0.53	0.40	0.48
	2	0.59	0.69	0.55	0.62
	3	0.68	0.77	0.64	0.69
	4	0.74	0.82	0.69	0.73
	5	0.78	0.85	0.73	0.75
6	1	0.68	0.74	0.60	0.65
	2	0.81	0.85	0.70	0.73
	3	0.87	0.90	0.74	0.76
	4	0.90	0.92	0.76	0.78
	5	0.92	0.94	0.77	0.79
		Reliability coefficients for %CWS			
		Relative decisions (G-coefficient)		**Absolute decisions (D-coefficient)**	
Grade	***n* probes**	**3 min**	**5 min**	**3 min**	**5 min**
3	1	0.39	0.42	0.36	0.38
	2	0.56	0.59	0.50	0.52
	3	0.66	0.68	0.57	0.59
	4	0.72	0.74	0.62	0.64
	5	0.76	0.78	0.65	0.67
6	1	0.66	0.68	0.61	0.63
	2	0.79	0.81	0.72	0.74
	3	0.85	0.86	0.77	0.79
	4	0.88	0.89	0.80	0.81
	5	0.90	0.91	0.82	0.83

rules for CBM-W in German – that the person evaluating has a relevant impact on the accuracy of the measurements.

Implications for the Use of Written Expression Curriculum-Based Measurement as a Screening and Progress Monitoring Tool

Conclusions for the use of CBM-W in practice can be drawn primarily from the D-studies. It should be noted that these only shed light on the aspect of reliability and must be supplemented for an overall conclusion by findings on the validity of the various scoring methods in different grades (McMaster and Espin, 2007; Romig et al., 2017). If we look only at the reliability results, we should distinguish between the use of CBM-W in the context of universal screenings and progress monitoring. These are two quite different tasks, but ideally, CBM-W should be suitable for both purposes (Payan et al., 2019).

Screenings whose goal is to identify the weakest writers in a group (Dunn, 2020) are typical contexts for relative decisions based on subjects' rankings. For this reason, G-coefficients are informative here if the group (e.g., class or students at the same level) rather than an external benchmark is used as a reference. It has already been shown in previous studies that the standard procedure, namely the collection of a single writing sample of 3 min, is not suitable to achieve sufficient reliability >0.80 (or even >0.90) (Keller-Margulis et al., 2021). Rather, depending on the grade level and scoring method, the evaluation of two to four 3-min writing samples is necessary for this purpose. Increasing the writing time leads in some constellations to the fact that fewer writing samples must be collected, but the total effort does not necessarily decrease. For example, in grade 3, relative reliability >0.80 is achieved with CWS by four 3-min samples (=12 min of writing time) or by three 5-min samples (=15 min of writing time). Accordingly, the feasibility and time-consuming nature of CBM-W as a universal screening tool is the main reason CBM-W is rarely implemented in practice (Payan et al., 2019). On the other hand, it must be stated that there are currently no alternatives for economical, reliable, and valid procedures to detect at-risk children in the area of writing in the context of universal screenings (Saddler and Asaro-Saddler, 2013). This underlines the need to understand more precisely the factors influencing the measurement accuracy of CBM-W and thus be able to optimize the procedure. Also, it should be reconsidered whether feasibility could be improved by reducing the frequency of screenings. It is recommended to conduct a writing screening three times a year with all students (Hosp et al., 2016; Traga Philippakos and FitzPatrick, 2018). However, Keller-Margulis et al. (2016a) found little within-year variance in student growth across different measurement points in the year in their study and therefore suggest limiting oneself to a single screening per year in the fall.

G-Theory provides an additional reliability coefficient in the form of the dependability coefficient. The D-coefficient focuses on the level of performance, regardless of rank. It is thus preferable for progress monitoring, in which students are compared with their performance over time (Fan and Hansmann, 2015). Concerning this intended use of CBM-W, we can conclude that the present analyses indicate that CBM-W is not sufficiently reliable – at least in German and in the grade levels studied – to be recommended for progress monitoring. For a single writing sample of 5 min duration, the highest D-coefficient in level 3 is 0.49 (TWW), and in grade 6 is 0.65 (CIWS) and fails to achieve a reliability >0.80. Even by using multiple writing samples – which would be impractical for weekly assessments anyway – only one case (%CWS in grade 6 with four measurements of 3 or 5 min each) succeeds in achieving sufficient reliability. This result is, in fact, disappointing, but it reflects well our initial observation.

Limitations and Future Research

Finally, some methodological aspects should be discussed, which can be optimized in future studies by simple modifications. Reference has already been made to assigning children's texts to the raters, which leads to difficulties in interpreting the results. Texts were distributed to raters class by class. As a result, the facet "rater" is mixed with the factor class, and it is impossible to separate both factors' influence. Puranik et al. (2014) found significant differences between classes in writing instruction and the amount of time students spent on school writing activities in a study of kindergarten classes. This was reflected in a high variation in spelling and writing skills at the class level. This study also raises the possibility of a substantial influence of the "class" level on student performance. In future studies, children's texts should not be presented to raters on a class-by-class basis but should be randomized. Moreover, Bloch and Norman (2012) point out that it is also problematic when the same rater is involved in multiple subject ratings because rater variance is confounded with subject variance. Thus, if G-theory is used in the context of CBM, where there are usually always multiple samples of student performance, then randomization between tests and raters should continue consistently so that different raters evaluate different samples of a child.

A second possibility for optimization concerns the facet time. Only by including this facet in the %CWS method could the interesting interaction between student and story starter be uncovered. To consider the writing time as a facet for the other scoring methods, a marking in the text (or the change of pens) would be necessary (Christ et al., 2005; Keller-Margulis et al., 2016a) after every minute of writing time. Consideration of time is also reasonable in future G-studies of CBM-W because we still know too little about at what grade level and how great an increase in writing time is beneficial and therefore indicated.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

JW led the team in data collection and participated in analyzing the writing probes. Both authors contributed to the conception and design of the study, performed all the analyses, and wrote the manuscript.

ACKNOWLEDGMENTS

We would like to thank Ralph Bloch for his indispensable support in using G-String.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2022.919756/full#supplementary-material

REFERENCES

Alamargot, D., and Chanquoy, L. (2001). *Through the Models of Writing*, 1st Edn. Dordrecht: Springer Netherlands.

Allen, A. A., Jung, P. -G., Poch, A. L., Brandes, D., Shin, J., Lembke, E. S., et al. (2019). Technical adequacy of curriculum-based measures in writing in grades 1–3. *Read. Writ. Q.* 33, 1–25. doi: 10.1080/10573569.2019.1689211

Bloch, R., and Norman, G. R. (2012). Generalizability theory for the perplexed: a practical introduction and guide: amee guide no. 68. *Med. Teach.* 34, 960–992. doi: 10.3109/0142159X.2012.703791

Bloch, R., and Norman, G. R. (2021). *G_String_VI: User Manual*. Hamilton, ON: McMaster University. doi: 10.1007/SpringerReference_28001

Bouwer, R., Béguin, A., Sanders, T., and van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Lang. Test.* 32, 83–100. doi: 10.1177/0265532214542994

Brennan, R. L. (2001). *Generalizability Theory. Statistics for Social and Behavioral Sciences Ser*. New York, NY: Springer, doi: 10.1007/978-1-4757-3456-0

Briesch, A. M., Swaminathan, H., Welsh, M., and Chafouleas, S. M. (2014). Generalizability theory: a practical guide to study design, implementation, and interpretation. *J. Sch. Psychol.* 52, 13–35. doi: 10.1016/j.jsp.2013.11.008

Campbell, H. M., Espin, C. A., and McMaster, K. L. (2013). The technical adequacy of curriculum-based writing measures with English learners. *Read. Writ.* 26, 431–452. doi: 10.1007/s11145-012-9375-6

Cardinet, J. (1998). Von der klassischen testtheorie zur generalisierbarkeitstheorie : der beitrag der varianzanalyse. *Bildungsforschung Und Bildungspraxis: Schweiz. Z. Erziehungswiss.* 20, 271–288.

Christ, T. J., and Hintze, J. M. (2007). "Psychometric considerations when evaluating response to intervention," in *Handbook of Response to Intervention: The Science and Practice of Assessment and Intervention*, eds S. R. Jimerson, M. K. Burns, and A. M. VanDerHeyden (Heidelberg: Springer), 93–105.

Christ, T. J., Johnson-Gros, K. N., and Hintze, J. M. (2005). An examination of alternate assessment durations when assessing multiple-skill computational fluency: the generalizability and dependability of curriculum-based outcomes within the context of educational decisions. *Psychol. Sch.* 42, 615–622. doi: 10.1002/pits.20107

Christ, T. J., Van Norman, E. R., and Nelson, P. M. (2016). "Foundations of fluency-based assessments in behavioral and psychometric paradigms," in *The Fluency Construct: Curriculum-Based Measurement Concepts and Applications*, eds K. D. Cummings and Y. Petscher (New York, NY: Springer), 143–163.

Deno, S. L. (1985). Curriculum-based measurement: the emerging alternative. *Except. Child.* 52, 219–232. doi: 10.1177/001440298505200303

Deno, S. L. (2003). Developments in curriculum-based measurement. *J. Spec. Educ.* 37, 184–192. doi: 10.1177/00224669030370030801

DESI-Konsortium (2006). *Unterricht und Kompetenzerwerb in Deutsch und Englisch*. Frankfurt: Zentrale Befunde der Studie Deutsch-Englisch-Schülerleistungen-International (DESI).

Dockrell, J. E., Connelly, V., Walter, K., and Critten, S. (2015). Assessing children's writing products: the role of curriculum based measures. *Br. Educ. Res. J.* 41, 575–595. doi: 10.1002/berj.3162

Dunn, M. (2020). "What are the Origins and Rationale for Tiered Intervention Programming?," in *Writing Instruction and Intervention for Struggling Writers: Multi-Tiered Systems of Support*, ed. M. Dunn (Newcastle upon Tyne: Cambridge Scholars Publisher), 1–15.

Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., and Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *J. Spec. Educ.* 34, 140–153. doi: 10.1037/spq0000138

Espin, C., Wallace, T., Campbell, H. M., Lembke, E. S., Long, J. D., and Ticha, R. (2008). Curriculum-based measurement in writing: predicting the success of high-school students on state standards tests. *Except. Child.* 74, 174–193. doi: 10.1177/001440290807400203

Fan, C.-H., and Hansmann, P. R. (2015). Applying generalizability theory for making quantitative RTI progress-monitoring decisions. *Assess. Effect. Interv.* 40, 205–215. doi: 10.1177/1534508415573299

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *Sch. Psychol. Rev.* 33, 188–193.

Fuchs, L. S. (2017). Curriculum-based measurement as the emerging alternative: three decades later. *Learn. Disabil. Res. Pract.* 32, 5–7. doi: 10.1111/ldrp.12127

Fuchs, L. S., and Fuchs, D. (2007). *Using CBM for Progress Monitoring in Written Expression and Spelling*. Available online at: https://files.eric.ed.gov/fulltext/ED519251.pdf (accessed March 31, 2022).

Fuchs, L. S., Deno, S. L., and Marston, D. (1983). Improving the reliability of curriculum-based measures of academic skills for psychoeducational decision making. *Diagnostique* 8, 135–149. doi: 10.1177/073724778300800301

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., and Slider, N. J. (2002). Moving beyond total words written: the reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *Sch. Psychol. Rev.* 31, 477–497.

Gansle, K. A., VanDerHeyden, A. M., Noell, G. H., Resetar, J. L., and Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *Sch. Psychol. Rev.* 35, 435–450.

Graham, S., Hebert, M., Paige Sandbank, M., and Harris, K. R. (2016). Assessing the writing achievement of young struggling writers. *Learn. Disabil. Q.* 39, 72–82. doi: 10.1177/0731948714555019

Graham, S., and Perin, D. (2007). *Writing Next: Effective Strategies to Improve Writing of Adolescents in Middle and High Schools – A report to Carnegie Corporation of New York*. Washington, DC: Alliance for Excellent Education.

Hintze, J. M., Owen, S. V., Shapiro, E. S., and Daly, E. J. (2000). Generalizability of oral reading fluency measures: application of G theory to curriculum-based measurement. *Sch. Psychol. Q.* 15, 52–68. doi: 10.1037/h0088778

Hooper, S. R., Swartz, C. W., Wakely, M. B., de Kruif, R. E. L., and Montgomery, J. W. (2002). Executive functions in elementary school children with and without problems in written expression. *J. Learn. Disabil.* 35, 57–68. doi: 10.1177/002221940203500105

Hosp, J. L., and Kaldenberg, E. (2020). "What is writing assessment for tiered decision making?," in *Writing Instruction and Intervention for Struggling Writers: Multi-Tiered Systems of Support*, ed. M. Dunn (Newcastle-upon-Tyne: Cambridge Scholars Publisher), 70–85.

Hosp, M. K., Hosp, J. L., and Howell, K. W. (2016). *The ABC's of CBM: A Practical Guide to Curriculum-Based Measurement. The Guilford Practical intervention in the Schools Series*, Second Edn. New York, NY: The Guilford Press.

Jewell, J., and Malecki, C. K. (2005). The utility of CBM written language indices: an investigation of production-dependent, production-independent, and accurate-production scores. *Sch. Psychol. Rev.* 34, 27–44.

Keller-Margulis, M. A., Mercer, S. H., and Matta, M. (2021). Validity of automated text evaluation tools for written-expression curriculum-based measurement:

a comparison study. *Read. Writ.* 34, 2461–2480. doi: 10.1007/s11145-021-10153-6

Keller-Margulis, M. A., Mercer, S. H., and Thomas, E. L. (2016a). Generalizability theory reliability of written expression curriculum-based measurement in universal screening. *Sch. Psychol. Q.* 31, 383–392. doi: 10.1037/spq0000126

Keller-Margulis, M. A., Payan, A., Jaspers, K. E., and Brewton, C. (2016b). Validity and diagnostic accuracy of written expression curriculum-based measurement for students with diverse language backgrounds. *Read. Writ. Q.* 32, 174–198. doi: 10.1080/10573569.2014.964352

Kent, S. C., and Wanzek, J. (2016). The relationship between component skills and writing quality and production across developmental levels. *Rev. Educ. Res.* 86, 570–601. doi: 10.3102/0034654315619491

Kim, Y. -S. G., Schatschneider, C., Wanzek, J., Gatlin, B., and Al Otaiba, S. (2017). Writing evaluation: rater and task effects on the reliability of writing scores for children in grades 3 and 4. *Read. Writ.* 30, 1287–1310. doi: 10.1007/s11145-017-9724-6

Malecki, C. K., and Jewell, J. (2003). Developmental, gender, and practical considerations in scoring curriculum-based measurement writing probes. *Psychol. Sch.* 40, 379–390. doi: 10.1002/pits.10096

McMaster, K. L., and Espin, C. (2007). Technical features of curriculum-based measurement in writing. *J. Spec. Educ.* 41, 68–84. doi: 10.1177/00224669070410020301

McMaster, K. L., Shin, J., Espin, C. A., Jung, P. -G., Wayman, M. M., and Deno, S. L. (2017). Monitoring elementary students' writing progress using curriculum-based measures: grade and gender differences. *Read. Writ.* 30, 2069–2091. doi: 10.1007/s11145-017-9766-9

National Center for Education Statistics (2011). *The Nation's Report Card: Writing 2011*. Available online at: https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf (accessed March 29, 2022).

Nunnally, J. C. (1967). *Psychometric Theory,* 5th Edn. New York, NY: McGraw-Hill.

Payan, A. M., Keller-Margulis, M. A., Burridge, A. B., McQuillin, S. D., and Hassett, K. S. (2019). Assessing teacher usability of written expression curriculum-based measurement. *Assess. Effect. Interv.* 45, 51–64. doi: 10.1177/1534508418781007

Poch, A. L., Allen, A. A., Jung, P.-G., Lembke, E. S., and McMaster, K. L. (2021). Using data-based instruction to support struggling elementary writers. *Interv. Sch. Clin.* 57, 147–155. doi: 10.1177/10534512211014835

Puranik, C. S., Al Otaiba, S., Sidler, J. F., and Greulich, L. (2014). Exploring the amount and type of writing instruction during language arts instruction in kindergarten classrooms. *Read. Writ.* 27, 213–236. doi: 10.1007/s11145-013-9441-8

Ritchey, K. D., McMaster, K. L., Al Otaiba, S., Puranik, C. S., Kim, Y. -S. G., Parker, D. C., et al. (2016). "Indicators of fluent writing in beginning writers," in *The Fluency Construct: Curriculum-Based Measurement Concepts and Applications*, eds K. D. Cummings and Y. Petscher (New York, NY: Springer), 21–66.

Romig, J. E., Miller, A. A., Therrien, W. J., and Lloyd, J. W. (2020). Meta-analysis of prompt and duration for curriculum-based measurement of written language. *Exceptionality* 29, 133–149. doi: 10.1080/09362835.2020.1743706

Romig, J. E., Therrien, W. J., and Lloyd, J. W. (2017). Meta-analysis of criterion validity for curriculum-based measurement in written language. *J. Spec. Educ.* 51, 72–82. doi: 10.1177/0022466916670637

Saddler, B., and Asaro-Saddler, K. (2013). Response to intervention in writing: a suggested framework for screening. Intervention, and Progress Monitoring. *Read. Writ. Q.* 29, 20–43. doi: 10.1080/10573569.2013.741945

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Lang. Test.* 22, 1–30. doi: 10.1191/0265532205lt295oa

Schoonen, R. (2012). "The validity and generalizability of writing scores: the effect of rater, task and language," in *Measuring Writing: Recent Insights into Theory, Methodology and Practices Studies in Writing*, Vol. 27, eds E. van Steendam, M. Tillema, G. Rijlaarsdam, and H. van den Bergh (Boston, MA: Brill), 1–22. doi: 10.5271/sjweh.3746

Stumpp, T., and Großmann, H. (2009). "Generalisierbarkeitstheorie," in *Enzyklopädie der Psychologie Methodologie und Methoden Evaluation: Bd. 1. Grundlagen und statistische Methoden der Evaluationsforschung*, eds H. Holling and N.-P. Birbaumer (Göttingen: Hogrefe Verl. für Psychologie), 207–234.

The National Commission on Writing in America's Schools and Colleges (2003). *The Neglected "R": The Need for a Writing Revolution*. Available online at: https://archive.nwp.org/cs/public/download/nwp_file/21478/the-neglected-r-college-board-nwp-report.pdf?x-r=pcfile_d (accessed March 29, 2022).

Traga Philippakos, Z. A., and FitzPatrick, E. (2018). A proposed tiered model of assessment in writing instruction: supporting all student-writers. *Insights Learn. Disabili.* 15, 149–173.

Troia, G. A., Shankland, R. K., and Wolbers, K. A. (2012). Motivation research in writing: theoretical and empirical considerations. *Read. Writ. Q.* 28, 5–28. doi: 10.1080/10573569.2012.632729

Weissenburger, J. W., and Espin, C. A. (2005). Curriculum-based measures of writing across grade levels. *J. Sch. Psychol.* 43, 153–169. doi: 10.1016/j.jsp.2005.03.002

Wilson, J., Chen, D., Sandbank, M. P., and Hebert, M. (2019). Generalizability of automated scores of writing quality in Grades 3–5. *J. Educ. Psychol.* 111, 619–640. doi: 10.1037/edu0000311

Winkes, J., and Schaller, P. (2022). Lernverlaufsdiagnostik schreiben (LVD – Schreiben): reliabilität, validität und sensitivität für mittelfristige lernfortschritte im deutschsprachigen raum. *Vierteljahress. Heilpädagogik Ihre Nachbargebiete* 91, 1–26. doi: 10.2378/vhn2022.art22d

Zheng, Y., and Yu, S. (2019). What has been assessed in writing and how? Empirical evidence from assessing writing (2000–2018). *Assess. Writ.* 42:100421. doi: 10.1016/j.asw.2019.100421

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.