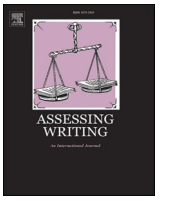
Assessing Writing 69 (2026) 101073

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw



Intraindividual variability in curriculum-based measurements of writing

Julia Winkes *

Department of Special Education, University of Fribourg, Fribourg, Switzerland

ARTICLE INFO

Keywords:
Curriculum-Based Measurement
Writing
Intraindividual Variability

ABSTRACT

This study investigates intraindividual variability (IIV) in curriculum-based measurements of writing (CBM-W) among primary school children. Inspired by the idea that such fluctuations may not only represent measurement error but also warrant description in relation to writing performance and development, the study examines patterns of IIV in this context. Data were collected from 345 children (51.6% female) in Grades 3–6 in Switzerland (mean age 10;5), including both monolinguals and multilinguals learning German. Students produced 10 writing samples at two time points (fall, spring), scored for correct writing sequences. The coefficient of variation quantified IIV at both times. Results show that IIV is systematically associated with performance level, with higher variability among lower-performing students. No clear age-related pattern was found once performance level was taken into account. IIV did not emerge as a predictor of subsequent writing development among students with comparable initial performance. In addition, students' home language background did not moderate the association between IIV and subsequent writing development. These findings suggest that IIV in CBM-W is closely related to performance level and highlight the importance of interpreting variability in relation to students' performance, calling for caution when using CBM-W in educational decision-making.

Curriculum-based measurement (CBM) is a core component of preventive support systems such as Response-to-Intervention and Multi-Tiered Systems of Support. Widely used for universal screening and progress monitoring, CBM enables teachers to make data-based decisions for both purposes through the regular use of short tests. (Silberglitt et al., 2016; Fuchs & Fuchs, 2017). Accordingly, the development of reliable, valid, and change-sensitive instruments that can be economically applied in daily school life is being pursued in various domains such as reading, math, or writing.

Despite careful test construction, children's performance often fluctuates from one assessment occasion to the next. Such fluctuations have implications both for the use of CBM at a single time point—for example, in the context of universal screening—and for its application in progress monitoring. At a single time point, high IIV raises concerns about the reliability of individual scores, as momentary fluctuations may obscure a student's typical level of performance (Salthouse, 2007). With respect to progress monitoring, variability can influence the interpretation of growth patterns; the presence of such fluctuations means that it takes a long time to reach a stable baseline or slope (O'Keeffe et al., 2017), and practitioners confronted with variable developmental data have legitimate difficulties interpreting the graphs clearly and basing their pedagogical decisions on them (Fan & Hansmann, 2015). To date, intraindividual variability (IIV) in the context of CBM has been attributed to measurement error and, consequently, discussed in relation to the reliability of the procedure. Ways of handling IIV in CBM have mainly involved aggregating data to get closer to the true score

* Correspondence to: Department of Special Education, University of Fribourg, Petrus-Kanisius-Gasse 21, Freiburg 1700, Switzerland.
E-mail address: julia.winkes@unifr.ch.

https://doi.org/10.1016/j.asw.2026.101073
Received 23 August 2025; Received in revised form 23 April 2026; Accepted 22 May 2026
Available online 1 June 2026

(Ford & Kaldenberg, 2019).

Fluctuations in human behavior —commonly described as within-person variation in performance across occasions or within a single session (MacDonald and Stawski, 2015)— are also a well-known phenomenon in other branches of research, and in some disciplines an expanded view of them has become established: IIV is regarded there as a distinctive and potentially meaningful feature of human behavior, representing an intrinsic aspect of cognitive processing (van Geert & van Dijk, 2002). From this point of view, the degree and pattern of IIV exhibited by an individual can carry important information and should be analyzed in addition to mean performance (Amenta & Crepaldi, 2016; Nesselroade & Molenaar, 2010). With regard to what such fluctuations indicate, two broad types of processes can be distinguished. One type arises from limitations or instability in core cognitive functions such as attention, working memory, and executive control. At this level, high IIV is generally considered a marker of vulnerability or processing difficulty and thus has negative implications for performance (Isbell et al., 2018; Cubillo et al., 2023). The other type of process is linked to active learning, such as adopting new strategies or integrating recently acquired knowledge. Here, temporary increases in variability can signal developmental change, as performance may become less stable while new skills are being consolidated (Lowie & Verspoor, 2019; Pfenninger & Kliesch, 2023).

What this means for IIV in children's writing development and for the potential significance of IIV in CBM-Writing (CBM-W) remains an open question. Writing is a cognitively demanding activity that relies heavily on working memory and executive control (Ruffini et al., 2024), capacities that are still developing in children. In addition, many component skills of writing—such as spelling, handwriting fluency, and planning—are not yet fully automated at this stage. As a result, young writers may be particularly susceptible both to performance fluctuations arising from cognitive resource limitations and to variability linked to the adoption of new strategies and skills.

In the following, we first outline the principles and procedures of CBM-W, the assessment format used in this study. We then examine IIV in CBM-W from two complementary angles. The first follows the classical test theory framework, which treats IIV primarily as measurement error and seeks to identify and minimize its sources. This perspective is particularly relevant for writing, since any consideration of IIV would be incomplete without first acknowledging its pronounced susceptibility to fluctuations among core academic skills. The second approach complements this view by considering that, alongside measurement error, part of the observed variability may reflect meaningful aspects of performance and development in writing. From this perspective, variability is not only a property of measurement but may also be understood as a characteristic of individual learners. In this sense, differences in performance stability between students become analytically relevant, raising the question of whether such fluctuations are purely unsystematic or systematically related to writing development. If such variability indeed carries meaning, then its occurrence in CBM-W data should first be examined systematically as a basis for evaluating its potential significance. The present study is an initial exploratory attempt to characterize intraindividual variability in CBM-W among children in Grades 3–6, focusing on its stability over time, its relationship to overall writing performance, and its potential role as a predictor of subsequent gains in writing proficiency. In this study, IIV refers to between-occasion variability based on two testing sessions conducted several months apart, each comprising ten writing samples collected over a two-week period. The association between IIV and writing performance was examined within each testing occasion, whereas stability and predictive relations were examined across the two occasions (T1 and T2).

1. Curriculum-based measurement of writing (CBM-W)

Curriculum-Based Measurement (CBM) was originally developed in special education to determine whether instructional programs were helping students progress toward their long-term goals (Deno, 2016). Early work by Deno (1985) established CBM as a technically adequate and efficient system for data-based decision making. Grounded in single-case logic, CBM relies on repeated measurement to evaluate students' responsiveness to instruction, as such designs are well suited for testing intervention effects (Deno, 2016). Standardized procedures for administering and scoring CBM across academic domains, including written expression, were subsequently developed (Deno, 1985; Romig et al., 2017). Building on this foundation, the basic principle of CBM in terms of progress monitoring is to use short, comparable learning samples repeatedly, to directly and longitudinally observe and graphically represent learning development. In addition, within a Response-to-Intervention (RTI) framework, CBM tasks are also widely used for universal screening to identify students who may be at risk for academic difficulties. Typically, tasks are used that integrate various sub-skills; thus, the results indicate the overall competency to be acquired, despite their brevity (Fuchs, 2017; Allen et al., 2019). This aligns with the idea of CBM as a general outcome measure (GOM), in which performance on brief, standardized tasks reflects broad curricular goals because multiple underlying skills are captured simultaneously (Silberglitt et al., 2016). Analogous to the domain of reading, writing fluency is used as an indicator of writing competence. It is operationalized by the amount of text that is written correctly and legibly within a certain period of time (Sturm et al., 2017). For students in the age range relevant to this study, a writing prompt is typically provided in the form of an introductory story starter (e.g., "Last week I was allowed to take my pet to school when…") and students are asked to write a response within a limited time period, usually between 3 and 7 min (Dockrell et al., 2015; Hosp & Kaldenberg, 2020). It should be noted, however, that story starters constitute only one of several established CBM-W task formats. Other validated formats—such as word- and sentence-level tasks—are also widely used in the field, particularly with younger students or those needing more foundational writing support (Ritchey et al., 2016). Various rating methods are available that quantify the amount and accuracy of the texts. These include the number of words written, the number of correct writing sequences (CWS), or the number of correct minus incorrect writing sequences. A writing sequence is defined as "two adjacent words that are correctly spelled, capitalized, and punctuated and that are grammatically and semantically acceptable to a native language speaker within the context of the sentence" (Saddler & Asaro-Saddler, 2013, p. 26). Consequently, with this method, each word is considered twice, once with the adjacent-before, and once with the adjacent-after. The duration of the writing sample and the choice of scoring method depend

primarily on the grade level. The consensus emerging across many studies is that longer writing samples and more complex scoring indices should be used as students get older (Payan et al., 2019; McMaster et al., 2017; Romig et al., 2017; Jewell & Malecki, 2005).

2. Intraindividual variability in CBM-W

2.1. Sources of intraindividual variability

When dealing with IIV, the first question that arises is which factors can lead to fluctuations in performance within an individual. In relation to school-based assessment, and particularly in the domain of writing, it is not surprising that students show variable performance across different test situations. The first source of variability is the *task* itself. Writing involves the integration of hierarchy-low (writing motor skills, spelling) and hierarchy-high processes (planning, formulating, revising). Writing is based, to a large extent, on experience with the task (e.g., knowledge of text schemata) and on working memory (Kim et al., 2017a; Ritchey et al., 2016). Reliably measuring such a complex activity poses difficulties; and Wilson et al. (2019), accordingly, question whether writing ability is sufficiently stable to allow for generalizable score inferences. In contrast to other school achievement tests, in writing it is not the assessment that determines the product to be evaluated (as is the case of a test with multiplication tasks or a word dictation); instead, the child produces the text to be evaluated on his or her own, and the same task evokes products that differ in many ways even among students with comparable writing skills (Ritchey et al., 2016). In addition, performance may vary due to child–task interactions, including cognitive and affective factors such as differences in background knowledge or motivation elicited by particular prompts (McMaster & Espin, 2007; Kim et al., 2017b).

The second possible source of intrapersonal variability is the *setting*, i.e. changes in internal and external conditions (van Dijk & van Geert, 2015). Bundock et al. (2018) discuss, for CBM Oral Reading Fluency, as setting factors the questions of *where* (e.g., classroom vs. individual quiet room), *by whom* (which administrator), and *how* (procedures used to administer assessments) the procedure was conducted.

The third possible source of variability in CBM-W is the *test* itself, and thus its reliability. Studies on the reliability of CBM-W have mainly investigated the influence of different story starters (parallel test reliability) and raters (inter-rater reliability). The agreement between different raters, in studies of CBM-W, is consistently in a high range, between 90% and 100% (Malecki & Jewell, 2003; Jewell & Malecki, 2005; Keller-Margulis et al., 2016b). However, this may also be due to the fact that raters often undergo extensive training in research (Allen et al., 2019). Specifically for CWS in the grade levels focused on here: parallel test reliability coefficients between.73 (Allen et al., 2019) and.79 (Weissenburger & Espin, 2005) are reported in English; while values between.54 (3rd Grade) and.74 (6th Grade) are reported in German (Winkes & Schaller, 2022b).

2.2. Magnitude of intraindividual variability

Parallel test reliability provides only incomplete insight into the equivalence of CBM parallel forms because it simply ascertains whether the ranking of individuals can be replicated (Ardoin & Christ, 2009; Fan & Hansmann, 2015). Specifically for the use of CBM in the context of progress monitoring, however, it is very important that there be comparability of the test scores at the individual-student level. In generalizability theory, this comparability is expressed through the D-coefficient, which reflects absolute (criterion-referenced) reliability—that is, how consistently an individual's observed score represents their actual level of performance across measurement conditions (Fan & Hansmann, 2015). Studies using generalizability theory to address the absolute reliability of CBM-W have not shown satisfactory results: a three-minute writing sample (correct minus incorrect writing sequences) achieved an absolute reliability of 0.48–0.61 in Keller-Margulis et al. (2016a). In Winkes & Schaller (2022b), a single writing sample, even with 5-minute writing time (CWS), achieved only D-coefficients between 0.45 and 0.59 depending on the grade level. This shows that repeated measurements with CBM-W within an individual do not lead to a stable assessment. From a decision-making perspective, such D-coefficients fall well below accepted standards for absolute reliability. High-stakes absolute decisions typically require values around.90, and even the more practical threshold of .80 is not reached. Thus, D-values in the .45–.59 range indicate that single CBM-W writing samples provide insufficient dependability for criterion-referenced decisions or for classifying students by proficiency levels (Kim et al., 2017b). Generalizability theory is also well-suited as a methodological framework to investigate the sources of variability. One interesting aspect can be highlighted in this respect: while it may seem plausible that different story starters induce variability between writing samples, this does not appear to be the case (Winkes & Schaller, 2022b). This reveals a difference from CBM Oral Reading Fluency where much of the variability between measurements could be reduced by careful construction of comparably difficult texts (Christ et al., 2016; Bundock et al., 2018). At the same time, generalizability-theory–based variance component analyses indicate that a substantial proportion of variance in CBM-W reflects within-student variability, with large residual components and, where modeled explicitly, person × task interactions accounting for roughly one third to more than half of the total variance, depending on grade level and scoring metric (Keller-Margulis et al., 2016a; Kim et al., 2017b). In a similar vein, longitudinal studies using CBM-W in progress-monitoring contexts have applied linear mixed-effects models and intraclass correlation coefficients to partition variance into within- and between-student components (e.g., McMaster et al., 2017; Valentine et al., 2021). These studies typically report ICCs indicating that approximately 30–40% of the total variance reflects within-student fluctuations over time, alongside substantial variability in individual intercepts and growth rates. Together, these methodological approaches provide detailed estimates of how much intraindividual variability occurs in CBM-W performance.

In summary, it can be noted that variability in children's academic performance—exemplified here by writing— is strongly influenced by the nature of the tasks and the context in which they are presented. Accordingly, these factors have been rightly

emphasized when optimizing testing conditions for CBM as progress-monitoring procedures, by constructing reliable procedures and eliminating confounding external influences. In the following section, this traditionally adopted view is expanded by considering that IIV may, beyond that, conceal more than just measurement error.

3. Intraindividual variability as meaningful data

A complementary research approach to that based on classical test theory and generalizability theory discussed above assumes that IIV– defined as "differences in behavior within the same individuals at different points in time" (Van Dijk and van Geert (2015, p. 37)– may be due to measurement error, but that fluctuations are also a characteristic feature of human behavior and are due to the dynamic and flexible architecture of cognitive processing (Amenta & Crepaldi, 2016). Following this perspective, variability bears important information; and IIV, consequently, becomes an object of research itself, instead of being averaged out using central tendency measures or data smoothing (Borella et al., 2011; van Geert & van Dijk, 2002; Nesselroade & Molenaar, 2010). The types of information hidden in IIV can be very diverse as Diehl, Hooker and Sliwinski (2015), XV) point out:

> Intraindividual variability rarely has a universal meaning but rather is indicative of different developmental processes at different developmental stages. For example, during childhood, greater intraindividual variability and particular patterns of intraindividual variability may reflect stages in learning and skill acquisition, whereas in later adulthood, similar patterns may reflect stages of adaptive transitions or increased vulnerability to decline.

Several methods are available to quantify IIV, the simplest being the individual standard deviation (Fagot et al., 2018). In order to control for differences in the mean, however, the coefficient of variation (CoV) is usually recommended. This is calculated individually for each person by dividing the standard deviation by the individual mean (Verspoor & de Bot, 2021; Huang et al., 2021; Borella et al., 2011). Van Dijk and van Geert (2015) note that, although variability can occur across different time scales, the concept is typically used to refer to short-term fluctuations spanning weeks, days, or even smaller units. The appropriate observation window, however, depends on the nature of the construct under study. More specific tasks require more frequent assessments, whereas for general skills that develop slowly (e.g., writing), Verspoor and de Bot (2021) consider 12 measurements across one year to be sufficient. Studies suggest that performance fluctuations represent a relatively stable characteristic of individuals, with some persons being generally more variable than others, which is why variability itself is often seen as trait-like (Ali et al., 2018; Lowie & Verspoor, 2019; Dirk & Schmiedek, 2016).

Research on the potential significance of IIV in the context of writing currently comes primarily from two fields of research: cognitive developmental psychology, and L2 research using complex dynamic systems theory (CDST). These differ from each other in various respects, for example with regard to: the tasks used (basic cognitive functions (reaction time, working memory) versus tasks that require skill acquisition and complex strategy changes); their research methodologies (focus on individual case studies vs. group designs); the target groups; and the explanatory theoretical mechanisms for IIV. Both perspectives and their relation to writing are briefly outlined as follows.

3.1. IIV in the tradition of cognitive psychology

Cognitive psychology has a long tradition of studying IIV in relation to specific, basic cognitive skills. These studies usually employ reaction time (Isbell et al. 2018) or working memory tasks (Judd et al., 2021; Dirk & Schmiedek, 2016). These latter tasks are often used as a proxy for school performance, because working memory is an important predictor of learning, IQ, and achievement (Blume et al., 2022; Dirk & Schmiedek, 2016). Meanwhile, IIV in reaction time tasks is thought to be an index of attentional fluctuations from task-relevant goals (Fagot et al., 2018; Isbell et al., 2018). Specifically for the domain of writing, the proxy assumption seems conclusive: various studies have shown a key role for executive functions, such as working memory, attention control, self-regulation, and mental flexibility, in children's writing performance (Cordeiro et al., 2020; Salas & Silvente, 2020). One finding from studies with school children on the IIV of basic cognitive functions is that younger children have, on average, a higher IIV than older children (Dirk & Schmiedek, 2016). Furthermore, a higher IIV in working memory is associated with worse academic performance, e.g., in mathematics and language (Blume et al., 2022), and prognostically with lower mathematical performance three years later (Judd et al., 2021). Isbell et al. (2018) show that IIV in response time is a predictor of cognitive performance in preschool and has both direct and indirect associations with the academic readiness of preschool children and their performance in first grade.

Increased IIV is also associated with various developmental disorders. First and foremost, IIV is consistently described as a typical deficit of persons with attention deficit hyperactivity disorder (Borella et al., 2011; Gooch et al., 2012; Kuntsi & Klein, 2012; Ali et al., 2018); however, various studies also point to connections with other disorders such as autism spectrum disorder (Geurts et al., 2008), fetal alcohol spectrum disorder (Ali et al., 2018), and dyslexia (Borella et al., 2011). Dirk and Schmiedek (2016) report that children with lower fluid intelligence show higher variability in working memory tasks.

Although causal inferences about functional cognitive mechanisms are not permissible from such results (Dirk & Schmiedek, 2016), it is reasonable to assume that elevated IIV in basic cognitive functions is a vulnerability factor that identifies lower achievers and children with developmental disabilities. IIV, in this line of research, is thus generally given a negative connotation as a marker of impairments in information processing (Borella et al., 2011) and predictor of cognitive dysfunction (Ali et al., 2018), which applies to research with older adults as well as children with learning disabilities. The latter is an important target group of progress-monitoring measures such as CBM within a Response-to-Intervention system.

3.2. IIV from a complex dynamic systems perspective

A second way in which IIV may be functionally relevant can be illustrated by Complex Dynamic Systems Theory (CDST). Introduced to the field of second language acquisition by Larsen-Freemann (1997), CDST has since become a significant source of inspiration in applied linguistics research. Writing, in particular, has frequently been examined from this perspective in L2 learning, which makes CDST a useful example for considering how phases of heightened variability can be associated with developmental transitions, such as the adoption of new strategies or the integration of recently acquired knowledge. A basic assumption of CDST is that language is a complex, adaptive system that evolves in a largely non-linear, dynamic and highly contextualized manner (Fogal, 2020). Language development is characterized by unstable and so-called transitional phases which alternate with more stable phases. Instabilities are an inherent feature of language development (van Geert & van Dijk, 2002), but this does not mean that they occur entirely at random or always uniformly over the course of an individual learner's development. IIV is increased in phases when the learner explores new behaviors and strategies, which leads in time to a less stable system in these phases, but then subsequently results in learning growth (Verspoor & de Bot, 2021). In contrast to the negative connotation of IIV in basic cognitive processes described above, proponents of a CDST perspective refer to variability in (second) language acquisition as "required byproduct of the learning process" (Lowie & Verspoor, 2019, p.202), "potential driving force of development, (…) harbinger of change" (van Geert & van Dijk, 2002, p. 342 f) and, above all, as a predictor of development.

It is important to emphasize that variability in CDST is not considered a *cause* of learning but rather a surface feature or symptom of a dynamic system. This indicates that new strategies or modes of behavior are being experimentally validated so that learning can occur (Huang et al., 2021; see also Verspoor & de Bot, 2021). The alternation between transitional phases and stable phases is characterized by unique developmental trajectories: not all learners in a group (e.g. a class) go through the phase alternations at the same time. Thus, it follows that IIV, in this research paradigm, is preferably studied in the context of case studies rather than group designs. CDST is not primarily interested in modelling mean development, but rather individual development trajectories (Bulté & Housen, 2020). Across different single-case studies, it has been found that periods of increased variability are indeed followed by higher learning progress (for a summary, including the specific research methodology, cf. Fogal 2020 or Bulté & Housen 2020). However, there are also some studies that have examined differences in IIV between different individuals in group studies. Of particular interest are the two studies: by Lowie and Verspoor (2019) with 22 young Dutch learners of English in secondary school; and subsequently, by Huang et al. (2021) with a sample of 22 Chinese adults learning English at the university level. Lowie and Verspoor (2019) document a significant correlation between proficiency gains over the academic year and the CoV of 22 writing samples, collected and holistically evaluated in the course of the year. Inspired by this result, Huang et al. (2021) largely replicated the design. The English writing proficiency of 22 L1 Chinese adults at the university level was measured with 12 texts scored holistically over one academic year, and multiple regression analyses were conducted to identify the predictors of L2 writing proficiency and gains in proficiency over time. Their results indicate that traditional factors such as motivation, language aptitude and working memory did not significantly predict final L2 writing proficiency. However, when the CoV was included in the regression models, it emerged as a significant predictor of both final L2 writing proficiency and the gains made during the study period. Higher variability in writing performance over time was, therefore, positively associated with greater proficiency gains, accounting for a substantial portion of the variance in outcomes. The results of these two studies thus raise the question of whether IIV is also a significant predictor in the writing development of children in their L1, or whether these mechanisms differ between L1 and L2 learners.

4. The present study

Although fluctuating performance is a well-known phenomenon for children, teachers and parents, there is still limited research on IIV in the context of school (Dirk & Schmiedek, 2016; Judd et al., 2021). In the field of writing, IIV has, to date, only been addressed in L2 research. The overall goal of the present study is therefore to provide an initial exploratory description of IIV in the context of CBM-W among monolingual and multilingual children in Grades 3–6. Whereas prior G-theory and mixed-effects studies have primarily quantified within-student variability at the sample level (e.g., via variance components or ICCs), IIV is conceptualized here as an individual-level characteristic that varies between students. As outlined above, intraindividual variability in children's writing may reflect both instability in underlying cognitive processes and ongoing language-related developmental change. From a cognitive perspective, higher variability is therefore expected to be associated with lower performance, reflecting instability in core processes such as attention, working memory, and executive control. At the same time, from a CDST perspective, increased variability may also occur during phases of learning and reorganization and thus be associated with subsequent development. These perspectives provide a conceptual framework for interpreting potential patterns of association, rather than constituting directly testable predictions within the present design, particularly in the case of CDST, which typically requires dense, idiographic time-series data. Accordingly, the present study focuses on whether variability is systematically related to performance level and developmental progress, examining observable patterns of association.

Research Question 1:

How stable is IIV in CBM-W over a 6-month interval?

On the one hand, IIV has been described as trait-like, which would suggest a relatively high degree of stability between measurement points. On the other hand, from a Complex Dynamic Systems Theory (CDST) perspective, IIV is expected to increase temporarily during periods of learning and reorganization, when new knowledge is integrated and new strategies are explored. Because such developmental phases are unlikely to occur at the same time for all learners, IIV is not expected to show high stability across individuals. Taken together, these considerations lead to the hypothesis of a moderate correlation between IIV at the two

measurement points.

Research Question 2:

What is the relationship between performance level and IIV?

a. Is there a correlation between performance level and IIV?
b. To what extent is this relationship independent of students' age?
c. How do low-performing students differ from average-performing writers in their IIV?

Regarding RQ 2a, we assume that there is a correlation between performance level and IIV, but the direction is unknown and could also depend on the age of the children. On the one hand, variability in basic cognitive functions has been identified as a characteristic of weak learners. In addition, Verspoor and de Bot (2021) highlight that variability is more likely to occur in very unstable systems, which leads them to expect that a group of beginners should be more variable than an advanced group of learners with stabilized language systems. On the other hand, variability is the only significant predictor of writing proficiency in the study by Huang et al. (2021) and positively associated with writing performance. If a negative association between IIV and performance is observed, however, the question arises whether this reflects differences in performance per se or is partly attributable to age-related developmental patterns, as research on intraindividual variability in basic cognitive functions has documented higher variability in younger children (Dirk & Schmiedek, 2016; Fagot et al., 2018). The question of whether low-performing writers exhibit more or less variability than their peers has practical implications for the application of CBM-W. Progress monitoring is typically not used with all students in a class but specifically with at-risk children. For this reason, we compare the 25% of the weakest writers who are considered at-risk children in the Response-to-Intervention approach, with the remaining children in the respective grade level. This comparison allows for an assessment of the practical magnitude of these differences, which is particularly relevant for interpreting variability in applied CBM-W contexts.

Research Question 3:

Is intraindividual variability at T1 associated with individual differences in writing development over six months, beyond initial performance and grade level, and is this relationship moderated by learners' home language background?

Lowie and Verspoor (2019), working with adolescent L2 learners in a longitudinal classroom context, and Huang et al. (2021), studying adult L2 writers, reported a positive association between intraindividual variability and development in second language writing, such that learners who exhibited higher variability tended to show greater improvements over time. It remains unclear, however, whether similar patterns can be observed in younger learners in the context of writing development, particularly when differences in initial performance and grade level are taken into account. Drawing on these findings, we expect a similar association, while acknowledging that its generalizability to younger learners and different developmental contexts remains an open question.

In addition, it is examined whether this relationship differs as a function of learners' home language background. As existing evidence is limited to L2 contexts, this analysis explores whether the association between intraindividual variability and development generalizes to both monolingual and multilingual learners or differs between these groups.

5. Methods

5.1. Participants

This study was conducted in the German-speaking part of Switzerland. A total of 354 students from Grades 3–6 participated. For the purposes of data analysis, only subjects who participated in both the first and second measurements were considered. As a result, nine students were excluded from the analyses, so that the total sample here consists of 345 primary school children. The participating classes originate from schools that cover both rural and urban areas. Although the total sample contains about the same number of boys as girls, this is not consistently the case in the individual grade levels (see Table 1). The four grade levels also differ in the proportion of students learning German as a second language, and of those receiving some type of special education support. This term covers all interventions that are not part of regular teaching (e.g. speech therapy, reduced learning goals, support by a special-needs teacher).

5.2. Instrument

CBM-W: The participating students wrote 10 writing samples at both measurement points (T1 = September, T2 = March). They

Table 1
Sample characteristics by grade.

	N	N Classes	Gender (% fem. – male)	Age in month (M, SD)	German as L2 (%)	Special Education service or German as L2 Service (%)
Grade 3	68	7	57.4 – 42.6	106 (6.2)	35.3	26.4
Grade 4	107	9	57.9 – 42.1	119 (6.7)	33.7	15.0
Grade 5	92	9	40.2 – 59.8	134 (7.6)	48.2	26.1
Grade 6	78	8	51.3 – 48.7	144 (6.1)	44.6	21.8
Total	345	33	51.6 – 48.4	126.2 (15.1)	40.2	21.7

were given an introductory story starter and then one minute of planning time to think of a story. The subsequent writing time was three minutes. The instruction was in accordance with Hosp et al. (2016). Story starters were selected to correspond to the real-life interests of primary school children, and not to evoke a one-word response. The scoring guidelines closely follow Anglo-American procedures (e.g., Hosp et al. 2016 or Fuchs & Fuchs 2007), but also considering the characteristics of the German language. For example, 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, the evaluation of orthographic correctness would be incomplete without the inclusion of necessary commas. In addition, all nouns are capitalized in German, and thus capitalization is more complex than in English.

The texts were scored using CWS. This is a commonly employed parameter that has been empirically shown to be reliable and valid for use at different grade levels (Romig et al., 2017; Dockrell et al., 2015). According to Jewell and Malecki (2005), CWS measures writing fluency as a production-dependent index, because it depends on how much a student writes. It also considers many components of correct writing such as spelling, grammar, punctuation and capitalization. In the present study, the average parallel test reliability of the writing probes using CWS varies between .54 (Grade 3) and .74 (Grade 6). There is a moderate correlation between CWS and the global teacher judgment of the children's writing competence (between .42 in Grade 3 and .64 in Grade 4) and also with a German translation of Subtests 6 and 7 from the Test of Written Language 4 (TOWL−4, Hammil and Larsen, 2009), where the correlation varies between .34 (Grade 3) and .76 (Grade 4). These data on the reliability and validity of CBM-W in German are roughly comparable to those reported in English-language studies (e.g. Allen et al., 2019; Gansle et al., 2006), although the low values for Grade 3 must be taken into account critically when interpreting the results. More detailed analyses of the psychometric properties of the German CBM-W tasks, including grade-specific reliability and validity estimates, are reported elsewhere (Winkes & Schaller, 2022a; in German).

5.3. Procedure

Participation in the study was conditional on the consent of the participants, parents, teacher, and school principal. The data collection occurred over two weeks in autumn (September) and two weeks in spring (March) during lessons. Over the course of these two school weeks, the children wrote a writing sample every day. If this was not possible on one day, the writing sample was made up the next day. Teachers were instructed to perform the tasks with the whole class, following the standardized instruction and sequence of story starters. The writing samples were scored by bachelor's and master's students in speech-language pathology and special education programs. Beforehand, all raters underwent extensive training until they were confident in using the scoring procedures. Doubtful cases were continuously discussed within the research team and documented in a written form. However, no formal inter- or intra-rater reliability estimates were calculated for the present dataset.

5.4. Data analysis

To operationalize IIV, the CoV is calculated for both the 10 samples of T1 and the 10 samples of T2. The CoV results from the individual standard deviation divided by the individual mean (Verspoor & de Bot, 2021).

Pearson's r-correlation between the IIV to T1 and to T2 is documented to test the stability of IIV over time (RQ 1). Furthermore, the CoV-scores of the students at T1 and T2 are also compared using T-tests for dependent samples, because the correlation can be high even if all participants move in the same direction. RQ 2 addresses the relationship between performance level and IIV. Pearson's correlations are also used here—namely between the mean of the CWS at one measurement time and the individual CoV at the same time (T1 or T2). This information is documented separately for the overall sample as well as for the individual grade levels. In addition, to examine whether this relationship is independent of age, a multiple regression analysis was conducted with IIV (CoV at T1) as the dependent variable and performance level (CWS at T1) and age (in months) as predictors.

To identify possible group differences between low-performing writers and students without writing difficulties, the sample is dichotomized within the individual grade levels, based on the mean of T1 (CWS), into the weakest 25% and children with scores above this cut-off. The CoV values of these groups (T1) are then compared using T-tests. RQ3 examines whether intraindividual variability (IIV) at T1 explains additional variance in writing performance at T2 beyond initial performance and grade level, and whether this relationship is moderated by students' home language background. To address this question, hierarchical linear regression analyses were conducted with writing performance at T2 as the dependent variable. In the first step, initial performance at T1 and grade level were entered as control variables. In the second step, IIV at T1 (operationalized as the coefficient of variation) was added to test whether it explained additional variance in subsequent writing performance.

To examine potential differences between learner groups, a second set of analyses was conducted in a subsample including only students with clearly defined home language backgrounds. This subsample comprised students who reported German as their sole home language (L1 German; $n = 171$) and those who reported another language as their sole home language (L2 German; $n = 132$), resulting in a total of $N = 303$. Students reporting both German and another language as home languages were excluded from this analysis to ensure a clear group comparison. In this subsample, home language background was entered as an additional predictor, and an interaction term between IIV and home language background was included in a final step to test for moderation effects.

6. Results

6.1. Descriptive results

Table 2 presents descriptive characteristics of the data for the key variables (CWS, CoV, writing growth between T1 and T2) by grade level.

As revealed in Table 2, the average achievement gain between fall and spring was about 6 CWS across all grade levels. Accordingly, a mixed ANOVA shows that there is no significant interaction between time (within subjects) and grade level (between subjects) regarding CWS ($F(3, 341) = 0.59$, $p = .617$, partial $\eta^2 = 0.005$). However, there is a significant main effect for time of measurement ($F(1, 341) = 582.61$, $p < .001$, partial $\eta^2 = .63$), and also a main effect for grade level ($F(3, 341) = 58.51$, $p < .001$, partial $\eta^2 = .34$).

However, since the initial scores of the children naturally differ between the grade levels, the relative increase in performance must also be considered (individual T2 minus T1, divided by the individual mean). In this respect, grade levels differ from each other ($F(3, 341) = 38.55$, $p < .001$, $\eta^2 = .23$), with students making greater progress in the lower grades versus their initial scores than in the higher grades.

With regard to the CoV, the mixed ANOVA reveals a statistically significant interaction between time and group ($F(3,341) = 8.04$, $p < .001$, partial $\eta^2 = .06$). An analysis of the simple main effect of grade level shows that grades differ both at T1 ($p < .001$) and at T2 ($p < .001$). According to the Tukey HSD, Grade 3 differs significantly from all other grade levels at T1 (every $p < .001$). There are no significant differences in the average CoV between Grades 4, 5 and 6 at T1. At T2, Grade 3 again differs significantly from all other grades (every $p < .001$), and there is also a significant difference between Grades 4 and 6 (.06, $p < .05$).

6.2. Stability of IIV over time

The correlation between IIV at T1 and at T2, each operationalized by the CoV of the ten respective writing samples, is $r = .62$ ($p < .01$). A T-test for dependent samples is used to evaluate whether the scores of the participants at T1 and T2 differ significantly from each other. Variability between the writing probes is significantly higher at T1 compared to T2, $t(344) = 6.78$, $p < .001$. The effect size is small, with $d = 0.36$. Since the participating students were first introduced to CBM-W at T1, the higher variability at the first measurement time could possibly be explained by a lack of routine. The next step is thus to recalculate the CoV for T1—this time without the first two story starters. Without story staters 1 and 2, the correlation between the variability to T1 and T2 is $r = .61$ ($p < .001$). The mean of the CoV is $M = 0.35$ ($SD = 0.17$) with all ten writing samples, and $M = 0.33$ ($SD = 0.17$) with only eight writing samples. The difference between the variability to T1 and T2 remains significant ($t(344) = 4.81$, $p < .001$, $d = .26$).

6.3. Relationship between performance level and variability

6.3.1. Correlation between performance level and IIV

Table 3 presents the correlation coefficients between the children's mean performance level at T1 and T2, and the IIV, measured at the same time respectively. These were done separately for the four grade levels and for the total sample. A very stable pattern can be seen, with a moderate negative correlation between writing skill and IIV in all grades ranging from $r = -.57$ to $-.66$ ($p < .01$). Higher writing proficiency is thus associated with lower variability between the individual tests at all grade levels. The correlation coefficients of all grade levels are almost identical, especially for T2.

To further examine whether this relationship is independent of age, a multiple regression analysis was conducted with IIV at T1 as the dependent variable and performance level (CWS at T1) and age (in months) as predictors. The overall model was significant, $F(2, 342) = 131.86$, $p < .001$, explaining a substantial proportion of variance in IIV ($R^2 = .44$). Performance level was a strong and significant predictor ($\beta = -.62$, $p < .001$), whereas age did not explain additional variance when performance was taken into account ($\beta = -.08$, $p = .068$). This indicates that the association between IIV and performance is not attributable to age-related differences.

6.3.2. Differences between low-performing students and higher-performing students

To illustrate the magnitude of differences in IIV between performance groups, the weakest 25% of children, and thus the target group of CBM-W in each grade, were identified and their CoV was compared with that of their peers. Table 4 shows substantial differences in IIV between lower- and higher-performing writers across all grade levels. Children with low writing fluency consistently

Table 2
Descriptive statistics (M (SD)) by grade level.

		CWS		CoV		Growth T1 – T2	
Grade	N	T1	T2	T1	T2	Absolute in CWS	Relative to the mean*
3	68	8.02 (4.18)	14.30 (7.12)	0.52 (0.19)	0.40 (0.20)	6.27 (4.40)	0.89 (0.65)
4	107	18.52 (8.46)	24.42 (10.01)	0.32 (0.15)	0.29 (0.11)	5.90 (4.76)	0.41 (0.36)
5	92	20.49 (9.05)	26.23 (10.88)	0.30 (0.13)	0.28 (0.11)	5.73 (4.87)	0.32 (0.27)
6	78	28.46 (11.68)	35.07 (12.85)	0.28 (0.10)	0.23 (0.11)	6.61 (4.41)	0.25 (0.20)
Total	345	19.22 (11.06)	25.32 (12.45)	0.35 (0.17)	0.30 (0.14)	6.10 (4.64)	0.44 (0.45)

* Growth relative to the mean: individual mean CWS T2 – individual mean CWS T1 / individual mean of CWS T1

Table 3
Intercorrelations among mean CWS and CoV at T1 and T2.

Grade	N	T1	T2
3	68	-.63	-.62
4	107	-.66	-.65
5	92	-.57	-.66
6	78	-.59	-.64
Total	345	-.66	-.66

Note: All correlations are significant at $p < .01$.

Table 4
Differences in CoV (T1) between Lower-Performing and Higher-Performing Writers.

	Lower-Performing (PR ≤ 25)			Higher-Performing (PR > 25)						
Grade	N	M	SD	N	M	SD	*df*	*t*	p	Cohen`s *d*
3	18	0.72	0.23	50	0.45	0.12	20.43	4.80	< .001	1.75
4	29	0.49	0.19	78	0.27	0.07	31.18	5.77	< .001	1.79
5	23	0.41	0.16	69	0.27	0.11	90	4.77	< .001	1.16
6	19	0.38	0.12	59	0.24	0.08	76	5.70	< .001	1.50

exhibit higher variability across CBM-W samples than their peers. The magnitude of these differences is large, with effect sizes ranging from d = 1.16–1.79 ($p < .001$), indicating strong practical relevance from Grades 3–6.

6.4. Intraindividual variability as a predictor of subsequent writing performance and its moderation by home language background

Hierarchical regression analyses were conducted to examine whether intraindividual variability (IIV) at T1 explained additional variance in writing performance at T2 beyond initial performance and grade level. In the full sample, the model including initial performance and grade level accounted for a substantial proportion of variance in writing performance at T2 ($R^2 = .86$, $p < .001$). Initial performance at T1 was a strong predictor of subsequent writing performance ($\beta = .94$, $p < .001$), whereas grade level did not explain additional variance in T2 performance beyond initial performance ($\beta = -.02$, $p = .34$). Adding IIV at T1 did not explain additional variance ($\Delta R^2 = .00$, $p = .54$), and IIV was not a significant predictor of writing performance at T2 (Table 5).

To examine potential moderation by home language background, the analyses were repeated in a subsample including only students with German as their sole home language (L1) and those with another language as their sole home language (L2). In this subsample, the inclusion of IIV did not improve model fit, and neither home language background nor the interaction between IIV and home language background contributed significantly to the model (all ps ≥.34; see Table 6). Overall, these findings indicate that intraindividual variability was not associated with subsequent writing performance when controlling for initial performance and grade level, and that this pattern did not differ between monolingual and multilingual learners.

7. Discussion

CBM procedures are used in many schools to illustrate learning progress accurately and sensitively. Excessive variability between the individual assessments of a child—as is often observed in practice—poses a significant problem, as it reduces both the reliability of slopes in progress-monitoring data and the interpretability of single time-point assessments used for screening purposes (McMaster & Espin, 2007). Previous research in the context of the evaluation of CBM tests has thus always been accompanied by an effort to reduce IIV, which is considered to be the result of changing environmental factors and measurement errors, as much as possible. In addition to these important efforts to ensure the reliability of CBM procedures, IIV can also be seen as an inherent characteristic of human behavior, which itself is an important source of information about psychological processes (Amenta & Crepaldi, 2016). In the context

Table 5
Hierarchical regression analysis predicting writing performance at T2 (full sample).

Predictor	B	SE	β	p
Step 1				
Initial performance (T1)	1.06	.03	.94	< .001
Grade level	-.28	.29	-.02	.34
R^2	.86			
Step 2				
Initial performance (T1)	1.05	.03	.93	< .001
Grade level	-.30	.29	-.03	.31
IIV (CoV at T1)	−1.17	1.92	-.02	.54
ΔR^2	.00			

Table 6
Hierarchical regression analysis predicting writing performance at T2 with moderation by home language background.

Predictor	B	SE	β	p
Step 1				
Initial performance (T1)	1.08	.03	.95	< .001
Grade level	-.44	.31	-.04	.15
R^2	.86			
Step 2				
IIV (CoV at T1)	−1.99	2.09	-.03	.34
ΔR^2	.00			
Step 3				
Home language (L1 vs. L2)	-.37	.57	-.02	.52
ΔR^2	.00			
Step 4				
IIV × Home Language	.38	3.27	.01	.91
ΔR^2	.00			

of CBM, such fluctuations become directly visible because CBM relies on short, repeated tests of equal difficulty. This provides a unique opportunity to investigate IIV in the context of academic skills, an area in which such analyses have rarely been conducted due to methodological challenges (Borella et al., 2011; Dirk & Schmiedek, 2016). Against this background, the present study set out to provide an initial exploratory description of intraindividual variability in CBM-W among school-aged children. Students in Grades 3–6 completed 10 CBM-W tests within a two-week period in the fall and spring of a school year, allowing us to identify basic patterns of variability and examine how these relate to performance level and later writing performance. There was a moderate stability of the IIV shown between the two measurement points (r = .62), which was significantly higher on average at T1 than at T2. This is consistent with the assumption that IIV can basically be regarded as trait-like. However, there might also be individually different phases of more/less variability. The following discussion focuses on the second and third research questions, which address the relationship between intraindividual variability, performance level, and subsequent development.

7.1. Intraindividual variability in relation to performance level

With regard to the second research question, the present findings indicate that intraindividual variability is systematically associated with performance level, with higher variability observed among lower-performing students. This pattern is also reflected in group comparisons: the target group of CBM-W as a progress-monitoring instrument, namely low-performing students, shows significantly higher variability than their higher-performing peers. The magnitude of these differences is substantial, with effect sizes (Cohen's d) ranging from 1.16 to 1.79. From research on intraindividual variability in basic cognitive functions, an additional age-related effect might have been expected, as variability has often been shown to decrease from childhood into adolescence as cognitive systems become more efficient (e.g., Fagot et al., 2018). However, such an age-related pattern was not observed in the present study once performance level was taken into account. Instead, intraindividual variability was more closely linked to differences in performance than to age per se.

From a theoretical perspective, the observed pattern aligns with accounts that describe intraindividual variability as being more pronounced in less stable systems, such as those of beginning or less proficient learners (Verspoor & de Bot, 2021). Research rooted in cognitive psychology has associated increased intraindividual variability with limitations or ongoing development in information processing, in particular executive functioning and working memory (Dirk & Schmiedek, 2016; Isbell et al., 2018), which are critical for the process of writing in children (Cordeiro et al., 2020; Salas & Silvente, 2020). In the context of writing development, lower-performing writers are still in the process of automatising fundamental skills such as spelling and transcription, which place higher demands on cognitive resources and may lead to less stable processing, thereby contributing to greater variability in observed writing performance. At this point, a shortcoming becomes apparent that has already been noted by several authors: theoretical models that link intraindividual variability to underlying mechanisms of learning and development are still limited, and Dirk and Schmiedek (2016) in particular warn against converting between-person findings into causal links and thus drawing direct conclusions about functional mechanisms (see also Fagot et al., 2018). Accordingly, the present study, like much of the existing research, remains restricted to the level of observed performance patterns, and statements about underlying cognitive mechanisms are therefore not warranted.

7.2. Intraindividual variability and subsequent writing performance

With regard to the third research question, the present study examined whether intraindividual variability at T1 is associated with subsequent writing performance when initial performance and grade level are taken into account. The results showed that intraindividual variability did not explain additional variance in later writing performance beyond initial performance, indicating that students with higher variability at T1 did not show greater gains when starting from comparable performance levels. This pattern was consistent across analyses and was not moderated by students' home language background.

The interpretation of intraindividual variability as a marker of developmental change is closely associated with CDST. From this perspective, increased variability is assumed to precede phases of accelerated learning, reflecting processes of reorganization and

restructuring. Testing this assumption, however, requires analyses at the intraindividual level based on dense time-series data that capture the temporal coupling between variability and development. The present study does not meet these requirements but is instead comparable to studies with adolescent and adult L2 learners (Lowie & Verspoor, 2019; Huang et al., 2021), which also draw on CDST while examining variability–development relations at the group level. These studies reported a positive association between intra-individual variability and subsequent development, with more variable learners showing greater progress over time. In contrast, the present findings indicate that intraindividual variability does not predict subsequent writing performance when students start from comparable initial levels. One possible explanation for this discrepancy can be ruled out: the relationship was not moderated by students' home language background. Accordingly, there is no indication that the functional relevance of intraindividual variability for writing development differs between L1 and L2 learners.

Two complementary explanations may account for the divergence between the present findings and previous research.

First, the functional meaning of intraindividual variability may depend on the developmental status of the writing system and the operationalisation of writing performance. In the present study, learners are still acquiring and automatising basic writing skills, and performance is strongly constrained by cognitive resources such as working memory and attentional control. Under these conditions, increased variability is likely to reflect performance instability. In contrast, the studies by Lowie and Verspoor (2019) and Huang et al. (2021), conducted with adolescent and young adult L2 learners, capture variability at later stages of development and interpret it as reflecting processes of restructuring and expansion. This difference is further reinforced by the use of productivity-based (CWS) versus holistic measures of writing, which may capture different aspects of performance and thus relate differently to development.

Second, the studies differ in how intraindividual variability is conceptualised and temporally framed. In the present study, variability is assessed within a short time window around a relatively stable performance level and is defined as deviation from an individual reference point. Importantly, it is operationalised prior to the developmental interval and independently of the outcome, allowing for a strict test of its predictive value. It should be noted, however, that CDST does not specify the temporal interval over which increased variability is expected to precede or accompany developmental gains, raising the possibility of a temporal mismatch between the timing of variability assessment and the developmental processes under investigation. In contrast, previous studies estimate variability across extended developmental periods (e.g., Huang et al., 2021; Lowie & Verspoor, 2019), during which performance levels themselves are changing. Consequently, these designs capture the co-occurrence of variability and development rather than separating variability as a predictor from subsequent developmental outcomes. Future research is needed to better understand how differences in developmental stage, task characteristics, and temporal framing may account for the discrepant findings across studies.

7.3. Practical implications

IIV in the context of CBM is not only an interesting and so far largely unexplored field of research, it also has concrete implications for practice. Toste et al. (2024), for example, identify variability in the data as one of the main difficulties in understanding and interpreting CBM progress-monitoring graphs, as teachers struggle to distinguish typically occurring fluctuations from intervention effects. These difficulties were more pronounced for graphs with higher variability. At the same time, variability also poses challenges when CBM is used for screening purposes, as decisions are often based on single measurements that may not reliably reflect a student's typical performance. This limitation is well documented in writing research, which has shown that a single writing sample is not a valid indicator of students' writing ability (Graham et al., 2011). In the context of CBM, however, this issue becomes particularly salient, as repeated measurements make fluctuations between assessment points directly visible.

A common recommendation in writing assessment is therefore to base decisions on multiple writing samples rather than on single observations, which is particularly relevant in the context of high-stakes decisions. This also applies to progress monitoring, where more reliable estimates of individual development require either a sufficient number of measurement points or multiple writing samples at each time point. At the same time, however, such approaches are time-intensive and often difficult to implement in everyday school contexts. Short-term fluctuations in CBM-W data should therefore be interpreted with particular caution and should not be taken as immediate indicators of instructional effects, especially when based on a limited number of observations (McMaster et al., 2011).

It appears particularly important to make a clear distinction between absolute and relative variability. The CoV, which is preferably used in research, relativizes the IIV of a person to their mean. However, the variability that becomes apparent to practitioners in their daily work with students (e.g., in the context of CBM) does not correspond to the CoV but rather to the individual standard deviation. In Grade 4 in this study, for example, the average individual standard deviation for CWS at T1 is 3.83 for the low-performing writers and 5.78 for their higher performing peers. Thus, at first glance, the fluctuations appear to be greater among the good writers. However, if we put them into proportion with the mean value (8.66 for the struggling writers and 22.18 for the good writers), the picture is reversed. The patterns presented in this study, such as the relationship between low performance and increased IIV, are therefore not apparent from a practical perspective, or they even seem to behave in exactly the opposite way. This description is also consistent with studies on CBM Oral Reading Fluency that show that IIV is greater in absolute terms (i.e., correct words/minute) in high-performing students than in the typical CBM target group (O'Keeffe et al., 2017). It follows from this observation that confidence intervals, as they exist for many standardized CBM materials, should ideally be reported separately for different performance groups (Bundock et al., 2018). The present findings further suggest that such differentiation is more appropriately based on performance level than on age or grade level.

Practitioners should also be sensitized to the fact that variability in writing performance is not merely a measurement problem but an inherent feature of the construct being assessed. Due to the open-ended and language-productive nature of writing tasks, sources of

variability such as topic familiarity, background knowledge, and momentary fluctuations in attention cannot be fully controlled and are therefore an integral part of observed performance. In this sense, "variable data are not necessarily inaccurate" (Bundock et al., 2018, p. 280), but reflect the conditions under which writing performance is produced. Practitioners who are familiar with CBM in more highly standardized domains such as reading or mathematics—where careful test construction reduces variability—may underestimate the extent to which variability is inherent to writing and should therefore adjust their expectations regarding the stability of CBM-W data accordingly when interpreting students' performance and making instructional decisions.

7.4. Limitations and future research directions

Several limitations restrict the generalizability of this study's results. Although all raters underwent training and there was ongoing feedback on open questions, there was no systematic control of intra- and inter-rater reliability. As a result, part of the observed intraindividual variability may reflect rater-related inconsistency rather than true fluctuations in students' writing performance. This is particularly relevant given that shorter and less structured texts produced by lower-performing writers may offer greater scope for interpretive variation in scoring. Such texts often contain fewer scorable units and more ambiguous or borderline cases, which may increase the influence of individual rater decisions on the assigned scores. As a result, even small differences in scoring judgments could contribute disproportionately to observed variability in these students' performance. Future studies should therefore carefully control and report rater reliability in order to disentangle measurement-related variability from substantive performance fluctuations. The sample is also not balanced between the different grades, e.g. regarding gender proportion or the proportion of children with German as L2. Comparisons between grades in terms of both performance and variability may have been influenced by these factors. Furthermore, writing 10 texts within a two-week time frame may have negatively affected motivation for some students, which may also be a cause of individual performance fluctuations.

The aim of this paper was to provide an initial exploratory description of CBM data with regard to the patterns of IIV in relation to academic performance. As Borella et al. (2011) point out, studies on IIV in academic performance have thus far been a desideratum. Dirk and Schmiedek (2016) explain this, among other things, with reference to methodological difficulties, namely the need for a large number of psychometrically comparable test items. This is where CBM could be a helpful tool, because the development of high-quality CBM tests has been advanced for decades. However, it can be questioned whether the domain of writing is the best testing ground for such an analysis. As explained in detail above, it is in the nature of writing that a reliable assessment is inevitably associated with great difficulties, which is also sufficiently documented in the literature (Wilson et al., 2019; Kim et al., 2017b). The existence of measurement error is not in doubt, including from the perspective of viewing IIV as a potential source of information (van Geert & van Dijk, 2002). Writing is certainly the academic skill, of all school performance tasks, most susceptible to measurement error-related fluctuations, which thus limits its validity for exploring the meaning of IIV. CBM tests in other domains, which have a high number of parallel tests of the same difficulty and have empirically proven to be reliable and sensitive to change, would probably be better suited to investigate the importance of IIV in academic performance tests. At the same time, the availability of CBM across multiple academic domains offers the opportunity to examine intraindividual variability at the individual level both across different academic skills (e.g., reading, writing, mathematics) and across different types of tasks. Examining IIV within the same individuals across basic cognitive functions, automatable academic skills (e.g., reading fluency), and cognitively demanding, language-productive tasks (e.g., writing) may therefore provide an important starting point for advancing theoretical models of learning and development that conceptualize IIV as a meaningful mechanism.

Beyond this, to better understand the possible relationship between phases of increased IIV and learning trajectories, monitoring should be performed more closely, e.g., via weekly tests. Most existing studies using generalizability theory or linear mixed-effects models have conceptualized intraindividual variability primarily as a sample-level characteristic, quantifying how much variability occurs on average within students. An important implication for future research and practice is to complement this perspective by considering IIV also as a characteristic of individual learners. The present study was suited to address this question from a social reference frame, showing that children who exhibit higher IIV tend to demonstrate lower writing performance. Designs that collect CBM data frequently over longer periods would additionally allow the use of an individual reference frame, making it possible to examine assumptions that could not be empirically tested here—namely, whether individual children indeed show phases of higher or lower IIV over time, and how such phases are temporally related to subsequent writing development.

CRediT authorship contribution statement

Julia Winkes: Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the author used ChatGPT (OpenAI) to support language editing and improve clarity and style of the text. After using this tool, the author carefully reviewed and revised all content and takes full responsibility for the content of the published article.

Declaration of Competing Interest

There are no conflicts of interest. The work is original and has not been submitted for publication elsewhere.

Data availability

Data will be made available on request.

References

Ali, S., Kerns, K. A., Mulligan, B. P., Olson, H. C., & Astley, S. J. (2018). An investigation of intra-individual variability in children with fetal alcohol spectrum disorder (FASD). *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence, 24*(5), 617–637. https://doi.org/10.1080/09297049.2017.1302579

Allen, A. A., Jung, P.-G., Poch, A. L., Brandes, D., Shin, J., Lembke, E. S., & McMaster, K. L. (2019). Technical adequacy of curriculum-based measures in writing in grades 1–3. *Reading & Writing Quarterly, 33*, 1–25. https://doi.org/10.1080/10573569.2019.1689211

Amenta, S., & Crepaldi, D. (2016). Editorial: The variable mind? How apparently inconsistent effects might inform model building. *Frontiers in Psychology, 7*, 1–2. https://doi.org/10.3389/fpsyg.2016.00185

Ardoin, S. P., & Christ, T. J. (2009). Curriculum-based measurement of oral reading: standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set. *School Psychology Review, 38*(2), 266–283.

Blume, F., Irmer, A., Dirk, J., & Schmiedek, F. (2022). Day-to-day variation in students' academic success: The role of self-regulation, working memory, and achievement goals. *Developmental Science, 25*(6), Article e13301. https://doi.org/10.1111/desc.13301

Borella, E., Chicherio, C., Re, A. M., Sensini, V., & Cornoldi, C. (2011). Increased intraindividual variability is a marker of ADHD but also of dyslexia: A study on handwriting. *Brain and Cognition, 77*, 33–39.

Bulté, B., & Housen, A. (2020). Chapter 9. A critical appraisal of the CDST approach to investigating linguistic complexity in L2 writing development. In G. G. Fogal, & M. H. Verspoor (Eds.), *Language Learning & Language Teaching. Complex Dynamic Systems Theory and L2 Writing Development, 54* pp. 207–238). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.54.09bul.

Bundock, K., O'Keeffe, B. V., Stokes, K., & Kladis, K. L. (2018). Strategies for minimizing variability in progress monitoring of oral reading fluency. *TEACHING Exceptional Children, 50*(5), 273–281. https://doi.org/10.1177/0040059918764097

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

Cordeiro, C., Limpo, T., Olive, T., & Castro, S. L. (2020). Do executive functions contribute to writing quality in beginning writers? A longitudinal study with second graders. *Reading and Writing, 33*(4), 813–833. https://doi.org/10.1007/s11145-019-09963-6

Cubillo, A., Hermes, H., Berger, E., Winkel, K., Schunk, D., Fehr, E., & Hare, T. A. (2023). Intra-individual variability in task performance after cognitive training is associated with long-term outcomes in children. *Developmental Science, 26*(1), Article e13252. https://doi.org/10.1111/desc.13252

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219–232. https://doi.org/10.1177/001440298505200303

Deno, S. L. (2016). Data-Based Decision-Making. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of Response to Intervention: The Science and Practice of Multi-Tiered Systems of Support (2nd ed. 2016* (pp. 9–28). Springer.

Diehl, M., Hooker, K., & Sliwinski, M. J. (Eds.). (2015). *Handbook of Intraindividual Variability Across the Life Span*. Routledge.

Dirk, J., & Schmiedek, F. (2016). Fluctuations in elementary school children's working memory performance in the school context. *Journal of Educational Psychology, 108*(5), 722–739. https://doi.org/10.1037/edu0000076

Dockrell, J. E., Connelly, V., Walter, K., & Critten, S. (2015). Assessing children's writing products: The role of curriculum-based measures. *British Educational Research Journal, 41*(4), 575–595. https://doi.org/10.1002/berj.3162

Fagot, D., Mella, N., Borella, E., Ghisletta, P., Lecerf, T., & Ribaupierre, A. de (2018). Intra-individual variability from a lifespan perspective: A comparison of latency and accuracy measures. *Journal of Intelligence, 6*(1). https://doi.org/10.3390/jintelligence6010016

Fan, C.-H., & Hansmann, P. R. (2015). Applying generalizability theory for making quantitative RTI progress-monitoring decisions. *Assessment for Effective Intervention, 40*(4), 205–215. https://doi.org/10.1177/1534508415573299

Fogal, G. G. (2020). Investigating variability in L2 development: Extending a complexity theory perspective on L2 writing studies and authorial voice. *Applied Linguistics, 41*(4), 575–600. https://doi.org/10.1093/applin/amz005

Ford, J. W., & Kaldenberg, E. R. (2019). Curriculum-based measurement for written expression with postsecondary students with intellectual and developmental disabilities. *Journal of Inclusive Postsecondary Education, 1*(2), 1–22. https://doi.org/10.13021/JIPE.2019.2473

Fuchs, L. S. (2017). Curriculum-based measurement as the emerging alternative: Three decades later. *Learning Disabilities Research & Practice, 32*(1), 5–7. https://doi.org/10.1111/ldrp.12127

Fuchs, L. S., & Fuchs, D. (2007). *Using CBM for Progress Monitoring in Written Expression and Spelling*. ⟨https://files.eric.ed.gov/fulltext/ED519251.pdf⟩.

Fuchs, D., & Fuchs, L. S. (2017). Critique of the national evaluation of response to intervention: A case for simpler frameworks. *Exceptional Children, 83*(3), 255–268. https://doi.org/10.1177/0014402917693580

Gansle, K. A., VanDerHeyden, A. M., Noell, G. H., Resetar, J. L., & Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *School Psychology Review, 35*(4), 435–450.

Geurts, H. M., Grasman, R. P. P. P., Verté, S., Oosterlaan, J., Roeyers, H., van Kammen, S. M., & Sergeant, J. A. (2008). Intra-individual variability in ADHD, autism spectrum disorders and Tourette's syndrome. *Neuropsychologia, 46*(13), 3030–3041. https://doi.org/10.1016/j.neuropsychologia.2008.06.013

Gooch, D., Snowling, M. J., & Hulme, C. (2012). Reaction time variability in children with ADHD symptoms and/or dyslexia. *Developmental Neuropsychology, 37*(5), 453–472. https://doi.org/10.1080/87565641.2011.650809

Graham, S., Harris, K., & Hebert, M. (2011). *Informing Writing: The Benefits of Formative Assessment: A Carnegie Corporation Time to Act report*. ⟨https://www.carnegie.org/publications/informing-writing-the-benefits-of-formative-assessment/⟩.

Hammil, D.D., & Larsen, S.C. (2009). Test of Written Language TOWL-4 (4th ed.). Pro-Ed.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). The ABC's of CBM: A Practical Guide to Curriculum-based Measurement (Second edition). *The Guilford practical intervention in the schools series.* The Guilford Press.

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

Huang, T., Steinkrauss, R., & Verspoor, M. (2021). Variability as predictor in L2 writing proficiency. *Journal of Second Language Writing, 52*, Article 100787. https://doi.org/10.1016/j.jslw.2020.100787

Isbell, E., Calkins, S. D., Swingler, M. M., & Leerkes, E. M. (2018). Attentional fluctuations in preschoolers: Direct and indirect relations with task accuracy, academic readiness, and school performance. *Journal of Experimental Child Psychology, 167*, 388–403. https://doi.org/10.1016/j.jecp.2017.11.013

Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review, 34*(1), 27–44.

Judd, N., Klingberg, T., & Sjöwall, D. (2021). Working memory capacity, variability, and response to intervention at age 6 and its association to inattention and mathematics age 9. *Cognitive Development, 58*, 1–9. https://doi.org/10.1016/j.cogdev.2021.101013

Keller-Margulis, M. A., Mercer, S. H., & Thomas, E. L. (2016a). Generalizability theory reliability of written expression curriculum-based measurement in universal screening. *School Psychology Quarterly: The Official Journal of the Division of School Psychology, American Psychological Association, 31*(3), 383–392. https://doi.org/10.1037/spq0000126

Keller-Margulis, M. A., Payan, A., Jaspers, K. E., & Brewton, C. (2016b). Validity and diagnostic accuracy of written expression curriculum-based measurement for students with diverse language backgrounds. *Reading & Writing Quarterly, 32*(2), 174–198. https://doi.org/10.1080/10573569.2014.964352

Kim, Y.-S. G., Gatlin, B., Al Otaiba, S., & Wanzek, J. (2017a). Theorization and an empirical investigation of the component-based and developmental text writing fluency construct. *Journal of Learning Disabilities, 51*(4), 1–16. https://doi.org/10.1177/0022219417712016

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

Kuntsi, J., & Klein, C. (2012). Intraindividual variability in ADHD and its implications for research of causal links. *Current Topics in Behavioral Neurosciences, 9*, 67–91. https://doi.org/10.1007/7854_2011_145

Larsen-Freemann, D. (1997). Chaos/complexity science and second language acquisition. *Applied Linguistics, 18*(2), 141–165. https://doi.org/10.1093/applin/18.2.141

Lowie, W. M., & Verspoor, M. H. (2019). Individual differences and the ergodicity problem. *Language Learning, 69*, 184–206. https://doi.org/10.1111/lang.12324

MacDonald, S. W., & Stawski, R. S. (2015). Intraindividual Variability - An Indicator of Vulnerability or Resilience in Adult Development and Aging? In M. Diehl, K. Hooker, & M. J. Sliwinski (Eds.), *Handbook of Intraindividual Variability Across the Life Span* (pp. 231–257). Routledge.

Malecki, C. K., & Jewell, J. (2003). Developmental, gender, and practical considerations in scoring curriculum-based measurement writing probes. *Psychology in the Schools, 40*(4), 379–390. https://doi.org/10.1002/pits.10096

McMaster, K. L., Du, X., Yeo, S., Deno, S. L., Parker, D., & Ellis, T. (2011). Curriculum-based measures of beginning writing: Technical features of the slope. *Exceptional Children, 77*(2), 185–206. https://doi.org/10.1177/001440291107700203

McMaster, K. L., & Espin, C. (2007). Technical features of curriculum-based measurement in writing. *The Journal of Special Education, 41*(2), 68–84. https://doi.org/10.1177/00224669070410020301

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

Nesselroade, J. R., & Molenaar, P. C. M. (2010). Emphasizing intraindividual variability in the study of development over the life span. In R. M. Lerner, M. E. Lamb, & A. M. Freund (Eds.), *The Handbook of Life-Span Development, 28* p. 1). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470880166.hlsd001002.

O'Keeffe, B. V., Bundock, K., Kladis, K. L., Yan, R., & Nelson, K. (2017). Variability in DIBELS next progress monitoring measures for students at risk for reading difficulties. *Remedial and Special Education, 38*(5), 272–283. https://doi.org/10.1177/0741932517713310

Payan, A. M., Keller-Margulis, M. A., Burridge, A. B., McQuillin, S. D., & Hassett, K. S. (2019). Assessing teacher usability of written expression curriculum-based measurement. *Assessment for Effective Intervention, 45*(1), 51–64. https://doi.org/10.1177/1534508418781007

Pfenninger, S. E., & Kliesch, M. (2023). Variability as a functional marker of second language development in older adult learners. *Studies in Second Language Acquisition, 45*(4), 1004–1030. https://doi.org/10.1017/S0272263123000013

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

Romig, J. E., Therrien, W. J., & Lloyd, J. W. (2017). Meta-analysis of criterion validity for curriculum-based measurement in written language. *The Journal of Special Education, 51*(2), 72–82. https://doi.org/10.1177/0022466916670637

Ruffini, C., Osmani, F., Martini, C., Giera, W.-K., & Pecini, C. (2024). The relationship between executive functions and writing in children: A systematic review. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence, 30*(1), 105–163. https://doi.org/10.1080/09297049.2023.2170998

Saddler, B., & Asaro-Saddler, K. (2013). Response to intervention in writing: A suggested framework for screening, intervention, and progress monitoring. *Reading & Writing Quarterly, 29*(1), 20–43. https://doi.org/10.1080/10573569.2013.741945

Salas, N., & Silvente, S. (2020). The role of executive functions and transcription skills in writing: A cross-sectional study across 7 years of schooling. *Reading and Writing, 33*(4), 877–905. https://doi.org/10.1007/s11145-019-09979-y

Salthouse, T. A. (2007). Implications of within-person variability in cognitive and neuropsychological functioning for the interpretation of change. *Neuropsychology, 21*(4), 401–411. https://doi.org/10.1037/0894-4105.21.4.401

Silberglitt, B., Parker, D., & Muyskens, P. (2016). Assessment: periodic assessment to monitor progress. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of Response to Intervention: The Science and Practice of Multi-Tiered Systems of Support (2nd ed* (pp. 271–291). Springer.

Sturm, A., Nänny, R., & Wyss, S. (2017). Entwicklung hierarchieniedriger Schreibprozesse. In M. Philipp (Ed.), *Handbuch: Schriftspracherwerb und weiterführendes Lesen und Schreiben* (pp. 84–104). Beltz Juventa.

Toste, J. R., Filderman, M. J., Clemens, N. H., & Fry, E. (2024). Graph out loud: Pre-service teachers' data decisions and interpretations of CBM progress graphs. *Journal of Learning Disabilities, 222194241231768*. https://doi.org/10.1177/00222194241231768

Valentine, K. A., Truckenmiller, A. J., Troia, G. A., & Aldridge, S. (2021). What is the nature of change in late elementary writing and are curriculum-based measures sensitive to that change? *Assessing Writing, 50*, Article 100567. https://doi.org/10.1016/j.asw.2021.100567

van Dijk, M., & van Geert, P. (2015). The nature and meaning of intraindividual variability in development in the early life span. In M. Diehl, K. Hooker, & M. J. Sliwinski (Eds.), *Handbook of Intraindividual Variability across the Life-span* (pp. 37–58). Routledge Taylor & Francis Group.

van Geert, P., & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior and Development, 25*(4), 340–374. https://doi.org/10.1016/S0163-6383(02)00140-6

Verspoor, M., & de Bot, K. (2021). Measures of variability in transitional phases in second language development. *International Review of Applied Linguistics in Language Teaching.* , Article 000010151520210026. https://doi.org/10.1515/iral-2021-0026

Weissenburger, J. W., & Espin, C. A. (2005). Curriculum-based measures of writing across grade levels. *Journal of School Psychology, 43*(2), 153–169. https://doi.org/10.1016/j.jsp.2005.03.002

Wilson, J., Chen, D., Sandbank, M. P., & Hebert, M. (2019). Generalizability of automated scores of writing quality in Grades 3–5. *Journal of Educational Psychology, 111*(4), 619–640. https://doi.org/10.1037/edu0000311

Winkes, J., & Schaller, P. (2022a). Generalizability of Written Expression Curriculum-Based-Measurement in the German Language: What Are the Major Sources of Variability? *Frontiers in Education, 7*, Article 919756. https://doi.org/10.3389/feduc.2022.919756

Winkes, J., & Schaller, P. (2022b). Lernverlaufsdiagnostik Schreiben (LVD – Schreiben): Reliabilität, Validität und Sensitivität für mittelfristige Lernfortschritte im deutschsprachigen Raum. *Vierteljahresschrift Für Heilpädagogik Und Ihre Nachbargebiete, 91*, 1–26. https://doi.org/10.2378/vhn2022.art22d

Julia Winkes is a lecturer in the Department of Special Education at the University of Fribourg, Switzerland. Her research interests include curriculum-based measurement, progress monitoring, writing, and language sample analysis.