Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01.04
Volume 14 (2022), No. 1, 61–88
CC BY-NC-SA 4.0 Waxmann 2022

Christin Vanauer, Sarah Chromik, Philipp Doebler,
& Jörg-Tobias Kuhn

Curriculum-Based Measurement of Basic Arithmetic Competence: Do Different Booklets Represent the Same Ability?

Abstract

An important prerequisite of progress monitoring as one source to support instructional decision-making is the existence of equivalent booklets. This study assesses this prerequisite with respect to a German elementary school math curriculum-based measurement instrument (LVD-M 2-4; Strathmann & Klauer, 2012). Every second week of a 19-weeks period, n = 108 third and n = 109 fourth graders (regular instruction) completed one of ten parallel booklets, each containing 24 arithmetic tasks. Analyses with (generalized) linear mixed models showed that in both grades the between-booklet variance was so small in relation to the between student variance that it was practically irrelevant. This corresponds to the key assumption of the binomial model that equivalent scores from different booklets reflect the same ability. While item difficulty varied within some of the tasks, the effect was insubstantial in comparison with the variance between students. These findings were replicated in two intervention samples of an RTI study. The parallel booklets can therefore be regarded as equivalent for typical applied purposes. Implications of these findings for curriculum-based measurement and booklet design are discussed.

Keywords

curriculum-based measurement, elementary school mathematics, formative assessment, booklet equivalence

Dr. Christin Vanauer (corresponding author) · Prof. Dr. Jörg-Tobias Kuhn, Faculty of Rehabilitation Sciences, TU Dortmund University, Emil-Figge-Str. 50, 44227 Dortmund, Germany
email: christin.vanauer@tu-dortmund.de
tobias.kuhn@tu-dortmund.de

Sarah Chromik, M. Sc. · Prof. Dr. Philipp Doebler, Faculty of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany
email: sarah.chromik@tu-dortmund.de
doebler@tu-dortmund.de

Christin Vanauer, Sarah Chromik, Philipp Doebler, & Jörg-Tobias Kuhn

Verlaufsdiagnostik arithmetischer Grundkompetenzen: Messen verschiedene Booklets die gleiche Fähigkeit?

Zusammenfassung

Eine wichtige Voraussetzung dafür, Lernverlaufsdiagnostik für instruktionale Entscheidungen nutzen zu können, sind äquivalente Testbooklets. Diese Studie prüft diese Voraussetzung für die „Lernverlaufsdiagnostik – Mathematik 2-4“ (LVD-M 2-4; Strathmann & Klauer, 2012). Über 19 Wochen hinweg bearbeiteten n = 108 Drittklässler:innen und n = 109 Viertklässler:innen in zweiwöchigem Abstand zehn verschiedene Paralleltests mit je 24 arithmetischen Aufgaben. Mithilfe (generalisierter) gemischter linearer Modelle wurden Booklet-Effekte in Relation zur Leistungsvarianz zwischen den Kindern gesetzt. Damit wurde die Kernannahme des Binomial-Modells geprüft, dass gleiche Scores aus verschiedenen zufallsgenerierten Booklets die gleiche latente Fähigkeit abbilden sollten. In beiden Klassenstufen fiel die Between-Booklet-Varianz in Relation zur Varianz zwischen den Kindern sehr gering aus. Für einige Aufgabentypen variierte die Schwierigkeit zwar zwischen den Booklets, war verglichen mit der Varianz der Schülerleistung aber nicht substanziell. Die Befunde ließen sich in zwei Interventionsgruppen einer RTI-Studie replizieren. Die Booklets können also für typische Anwendungszwecke als äquivalent angesehen werden. Die Implikationen dieser Befunde werden vor dem Hintergrund von Lernverlaufsdiagnostik und der Konstruktion äquivalenter Testbooklets diskutiert.

Schlagworte

Lernverlaufsdiagnostik, Mathematik in der Grundschule, formatives Assessment, Booklet-Äquivalenz

1. Theoretical Background

1.1 Standardized Diagnostics as a Contribution to Instructional Decision-Making

Instructional decision-making in schools is complex. Educational psychology can assist teachers by providing and evaluating instruments to systematically assess clear-cut competences of their students. While test construction in the domains of reading and writing faces a longer tradition, mathematical learning has mainly been targeted in the last two decades. In this period, a considerable number of tests have been published to help assess mathematical achievement in general, but also to identify children at risk for dyscalculia. Test backgrounds range from neuropsychological or developmental models to curricular or even higher-order national and international standards (Köller & Reiss, 2013; Kuhn & Schwenk, 2018).

Independent of the domain two main categories of standardized diagnostics can be distinguished: summative and formative approaches (Klauer, 2014), which differ in their purpose and measurement frequency. Summative tests (status diagnostics) evaluate a student's performance at a given point in time and help to identify individuals beyond a clinically relevant threshold. Unless designed as a screening, most status diagnostics are time-consuming and not suitable for large groups in everyday school practice. By contrast, formative testing, or progress monitoring, is based on multiple measurements that usually occur repeatedly (e.g., weekly). In contrast to summative testing, progress monitoring captures learning *courses* or trajectories (Deno, 2003). Progress monitoring is applicable to any competence on which subjects are expected to make visible progress in a reasonable period of time. It is described and recommended for mainly elementary school reading, spelling, and mathematics (especially basic arithmetic operations), but also for precursor abilities like letter naming or number identification (Foegen et al., 2007; Hosp et al., 2016).

1.2 Progress Monitoring in Special Needs Education and Beyond

Progress monitoring has gained particular importance in special needs education, for example, being a core feature of the *response to intervention* (RTI) paradigm, a preventive and inclusive educational concept with roots in the USA and transfer to other countries (e.g., Germany: Huber & Grosche, 2012; Voß, 2016, or Finland: Björn et al., 2018). In RTI, formative assessment helps to dynamically keep track of key learning criteria and thereby assign students to one of three tiers of educational support between regular classroom instruction (Level 1) and intensive single-subject fostering (Level 3). In frameworks like RTI, two things are important to bear in mind regarding the informative value of progress-monitoring tests: First, aggregated progress-monitoring outcomes (e.g., number of correctly read words in 1 minute or number of correctly solved calculations in a set of 24 tasks) are not sufficiently fine-grained to unravel the educational needs of low-performing students. In this case, class-level progress monitoring can only serve as a screening tool, painting the "big picture". Qualitative diagnostics like error analysis (e.g, Ashlock, 2005; Gerster, 2012) and/or the "thinking aloud" method (e.g., Lawson & Rice, 1987) are needed to find out where a specific student struggles. Second, progress monitoring is not restricted to this subgroup with special educational needs: It potentially informs data-based instruction of students in the whole ability range, even of very high achievers (Hebbecker & Souvignier, 2016). Therefore, research on the construction and analysis of progress-monitoring assessments is relevant for all types of schools and educational practice.

1.3 Particularities of Progress-Monitoring Tests

Progress-monitoring tests must meet certain requirements which are subject of research. To systematize research on curriculum-based measurement (CBM), Lynn Fuchs (2004) describes three stages. The first stage focuses on the psychometric properties (reliability, validity) of a progress-monitoring instrument and is not much different from the evaluation of status tests. However, the question of how to generate the required *set* of equivalent test booklets is not trivial. Several psychometric challenges are specifically linked to *progress* diagnostics: For instance, the "difficulty" of a specific item as well as the difficulty of the entire test has to decrease over time if learning progress is made (Rohwer, 2015). Simply speaking, a student is expected to make fewer errors over time in case that items are very similar. Moreover, test takers' ability is likely to develop differently in between (with inter-individually different growth curves). Most importantly, as a matter of validity, an increasing test score across time should not be due to rote learning of specific, repeatedly administered items but due to an increasing mastery of the competence represented by them.

This leads to the methodologically crucial second research stage outlined by Fuchs (2004). Studies in this stage assess whether a given instrument is in fact able to measure the core construct learning *progress* (also called "sensitivity to change"; Klauer & Strathmann, 2013). In addition to performance variability across measurement points, the validity of the learning slope as a predictor of learning progress, or as a criterion for response to intervention, is investigated. Most of respective studies are based on reading (Schatschneider et al., 2008; Stage & Jacobsen, 2001), only few on mathematical instruments, and they show heterogeneous results: Some studies report substantial correlations between learning progress and later achievement (Keller-Margulis et al., 2008), others do not find any or only negligible incremental validity for learning slopes (i.e., progress lines) beyond baseline ability (e.g., Shapiro et al., 2015). This inconclusive pattern could be partially related to the lack of booklet equivalence: an important prerequisite for data-based decision-making, which is addressed in third-stage research according to Fuchs (2004).

1.4 Construction of Progress-Monitoring Instruments

To overcome psychometric dilemmas linked to progress monitoring, the framework of generic test construction (Rohwer, 2015) defines performance based on correctly solved representations of task types (e.g., mental addition as a category of similar problems), each consisting of a pool of structurally similar (and thus theoretically equivalent) items. Hence, a new test booklet generated by randomly sampling an item of each task type.

Because the idea that the *same* test needs to be repeated is more a theoretical (i.e., content validity of the tasks types) than a technical (i.e., psychometric) matter in the first place, it is important that task types represent a meaningful standard. In the context of regular schools and average student populations, the curriculum is an obvious standard, as it defines the learning goals of each grade. In line with this, learning progress monitoring is predominantly thought of as CBM since emerging in the early 1970s (Deno, 1985; Klauer, 2014). Deducing appropriate task types from curricula is referred to as curriculum-sampling. Applying this approach to elementary school mathematics, the tasks included in CBM represent the set of, for example, arithmetic problem types that children of a specific grade should master at the end of the school year. Mathematical CBMs known to the authors of this study mainly focus on arithmetic operations (addition, subtraction, multiplication, division), but there are also examples comprising other curricular mathematical domains like geometry (e.g., quop; Souvignier, 2018).

However, curriculum-sampling is just one possibility, next to concentrating on so-called robust indicators (Fuchs, 2004; Schwenk et al., 2017). Robust indicators are key competences that are empirically valid for the achievement development in a specific domain and might be more basic than the curricular goals for a given grade. The main advantage of this approach is its flexibility: Robust indicator tests are seamlessly applicable across grade boundaries. While passage reading fluency is a well-established robust indicator in reading, there is no equivalent in mathematics. Foegen (2007) explains this with the differences of curricula, with math, compared to reading, learning being more complex and better represented by curriculum-sampling instead of robust indicators.

1.5 Booklet Equivalence

A key condition to interpret learning trajectories unambiguously is the assumption of equivalent test forms. Only when different test forms are equivalent can longitudinal test scores be considered as valid measures of individual learning. However, the degree to which different booklets of progress-monitoring instruments are equivalent or influenced by design features has only been addressed explicitly by few studies, mainly in the field of reading (CBM-R): Absolute estimates of weekly growth rates and intercepts as well as standard errors were shown to vary across different texts (“passages”) used for oral reading CBM (Ardoin & Christ, 2009; Francis et al., 2008). Therefore, absolute scores in CBM, that is, words read correctly per minute, should be interpreted with caution (Ardoin et al., 2013). By contrast, relative estimates of growth rates, that is rank orders of students, are more reliable.

Research in the field of mathematics (CBM-M) is comparably scarce (Montague et al., 2010). Like in CBM-R, Christ and Vining (2006) describe an “excessive reliance on research that has examined relative score interpretation” (p. 398). Stratified instead of randomly ordered booklets, that is, a presentation of items structured by the skill they assess (Christ & Vining, 2006), seem to enhance booklet equivalence.

How can cross-booklet equivalence be tested from a technical point of view? In the literature, the concept of equivalence has been investigated from varying perspectives, using at least three different yet related methodological approaches. First, in classical test theory, which operates at the level of test scores, two test forms are considered as parallel in the case that their mean test scores and test score variances are equal (e.g., McDonald, 1999). In the case of differing means and/or variances, equivalence can be established using test equating (e.g., Kolen & Brennan, 2014). In this approach, item parameters are of secondary importance. For example, Strathmann and Klauer (2010) compared means of adjacent math progress-monitoring test scores at each measurement point.

In contrast, item response theory (IRT) is a second approach that allows investigating test equivalence at the item level by analyzing differential item functioning (DIF). If no DIF exists between two test forms (i.e., if item parameters are identical), the test forms can be considered equivalent. Taking a less restrictive stance, equivalence of test forms within an IRT framework can also be assumed in the case that test information functions (TIF) are similar (Förster & Kuhn, 2021).

A third approach of analyzing test form equivalence relates to investigating the relative proportion of test score variance that can be attributed to test forms, in contrast to other factors (e.g., students, testing occasions). This is in the tradition of generalizability theory (e.g., Brennan, 2001). In this approach, the equivalence of test forms can be expressed in a relative way: The degree of equivalence is high if the variance attributable to test forms is very small, compared to the variance related to other factors. For example, Fan and Hansmann (2015) investigated a CBM of oral reading fluency, and found that 2.8% of total variance was attributable to probe variability, whereas 90.2% of variance was due to student reading skill. In the present study, we investigated test equivalence using this approach. In case the variance is solely attributable to student ability and unspecific error, but not to items/item families, there is neither DIF nor differential test functioning (DTF; i.e., when discussing Rasch models). In this sense, this is a strict approach, as the global absence of item (family) level variance implies equivalence in the IRT tradition and hence the classical test theory (CTT) tradition.

1.6 Summary: Aim of this Study

The key prerequisite of CBM instruments – a valid pool of different but psychometrically equivalent booklets – is often theoretically taken for granted. This is coupled with the assumption that identically distributed test scores stemming from different booklets are equivalent, which is part of a binomial model (cf. Klauer, 2011). Within the binomial model and under the prerequisite of random (stratified) item sampling, an individual's ability equals the proportion of correctly solved items on a power test, that is, accuracy.

Against this background, the aim of the present study is to test the booklet equivalence assumption of the binomial model empirically. Based on a curriculum-based measurement in mathematics, more specifically: basic arithmetic (LVD-M 2-4; Strathmann & Klauer, 2012), it takes a closer look at differences between booklets as one possible source to explain the variance of third and fourth graders' scores.

2. Method

2.1 Sample

The sample of this study is part of a larger research project funded by the German Federal Ministry of Education and Research (BMBF) and approved by a local ethics committee, with the aim to evaluate the effects of dyscalculia interventions within an RTI framework. A total of n = 687 students (n = 345 Grade 3, n = 342 Grade 4) of ten elementary schools in the same German region with urban and rural parts participated in an initial screening at the beginning of school year 2015/16.

After screening, schools were assigned to one of three experimental groups of roughly the same size (> 100 per group in each of the two Grades 3 and 4): a waiting control group and two intervention groups who completed bi-weekly basic arithmetic progress-monitoring tests throughout a 19-weeks phase (more closely described in Section 2.3). During this study period, all children of the waiting control group received regular classroom instruction, irrespective of their performance in the initial screening. In the two intervention groups, those children with at-risk level performance or below (PR ≤ 25) on the initial screening (arithmetic subscale of a grade-specific German mathematics test) received trainings according to different intervention schemes: One group followed a two-tiered schemed with computer-based training (Kuhn & Holling, 2014) for all children with at- or below-risk level performance and a three-tiered group with additional within-person small group training for the weakest 10% according to the initial arithmetic screening.

As the research aim of this study applies to regular school practice, we focus on the descriptions and results of the regularly instructed waiting control group: Children who did not complete any of the ten progress-monitoring tests (cf. Section 2.3) were excluded, leading to a final (waiting control) sample of n = 108 children in Grade 3 and n = 109 in Grade 4, with eighteen classrooms of four schools (Table 1). The number of participating children per classroom (i.e., children with parental consent obtained before participation) ranged between 6 and 24 (M = 12.06, SD = 4.78, median = 11). Next to mathematical ability measured with the arithmetic subtests of DEMAT 2+ (Krajewski et al., 2004) and DEMAT 3+ (Roick et al.,

2004) as part of the screening, general intelligence (CFT 1-R; Weiß & Osterland, 2013, or CFT 20-R; Weiß, 2006) and reading speed (SLS; Wimmer & Mayringer, 2014) were captured for all children participating in the study (Table 1).

Table 1: Sample (Waiting Control Group)

School	Grade 3 (*n* = 108, 58.3% male)				Grade 4 (*n* = 109, 55.0% male)			
	A	B	C	D	A	B	C	D
Students	26	23	40	19	21	16	48	24
Classrooms	2	3	2	2	2	2	3	2
Abilities	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max
Intelligence (CFT, IQ)	101.96	13.12	67.0	138.0	107.94[b]	16.52	61.0	148.0
Reading (SLS, RQ)	99.12[a]	14.11	65.1	134.9	98.03[b]	14.43	65.1	130.81
Arithmetic (DEMAT, *T*)	50.88[a]	8.87	28.0	71.0	49.07[c]	9.34	27.0	73.0

Note. CFT = Culture Fair Intelligence Test; IQ = intelligence quotient; SLS = Salzburger Lese-Screening [Salzburg reading screening test]; RQ = reading quotient; DEMAT = Deutscher Mathematiktest [German mathematics test].
The sample consists of the waiting control group of a larger response to intervention study. The four schools participating in this group are referred to as "A", "B", "C" and "D". IQ and RQ values approximately follow an *N*(100, 15) distribution, *T* values an *N*(50, 10) distribution. Some statistics have a lower sample size due to missing values: [a]: *n* = 107; [b]: *n* = 108; [c]: *n* = 106.

2.2 Progress-Monitoring Instrument LVD-M 2-4

The progress-monitoring instrument "Learning-progress diagnostics – mathematics for grades two to four [German original: Lernverlaufsdiagnostik – Mathematik für zweite bis vierte Klassen]" (LVD-M 2-4; Strathmann & Klauer, 2012) is applicable in the respective grades of elementary school. Each LVD-M 2-4 booklet contains 24 tasks that cover the grade's arithmetic (addition, subtraction, multiplication, division) part of the curriculum – in a way that it applies to all (or at least most) German states with their federal curricula.

While Grade 3 booklets (Table 2) consist of 19 mental (addition, subtraction, multiplication, division) and five written arithmetical problems (addition and subtraction), Grade 4 booklets (Table 3) mainly contain written tasks (18 items of all four basic arithmetic operations) and only six mental addition and subtraction items at the beginning. Beyond the solution mode (mental or written) and the basic arithmetic operation, the items also differ with regard to the task structure (Do the test takers have to calculate the result of an equation, like in 926 + 53 = ?, or fill in a blank, like in 874 + ? = 900), number range (place value structure), and the need to perform carry operations (e.g., tens-carry).

Table 2: Task Structure of the LVD-M 2-4 Booklets, Grade 3

Item(s)	Task structure	Place value structure/arithmetic operator	Mode	Example
1–2	a + b = ?	HTO + TO (with T+T ≤ 100)	m	926 + 53 = ?
3	c – b = ?	HTO – O (with ? > H, first O < second O)	m	982 – 3 = ?
4	c – b = ?	HTO – HTO	m	856 – 117 = ?
5	a + b = ?	HTO + TO (with ? ≤ 1000, TO + TO ≤ 100)	m	542 + 16 = ?
6	a + ? = H00*	HTO + TO	m	874 + ? = 900
7	c – ? = a	HTO – TO	m	967 – ? = 952
8, 9, 11	a · b = ?	times 6; times 9; times 50	m	4 · 6 = ?; ..., 7 · 50 = ?
10	a · ? = c	times 7	m	9 · ? = 63
12, 13	c = ? · b	times 8; times 20	m	24 = 8 · ?; 80 = 20 · ?
14–19	c : b = ?	divided by 6; ...by 9; ... by 8; ... by 7; ... by 20; ... by 50	m	24 : 6 = ?, ..., 200 : 50 = ?
20	a + b + c = ?	TO + TO + TO (with O + O + O > T, T + T + T > H)	w	15 + 95 + 39 =
21	a + b =?	HTO + HTO (with O + O > T, T + T < H, H + H < Th)	w	338 + 336 = ?
22	c – b = ?	HTO – HTO	w	876 – 741 = ?
23	c – b = ?	H00 – HTO	w	700 – 168 = ?
24	c – b = ?	1000 – HTO	w	1000– 439 = ?

Note. Th = thousand(s); H = hundred(s); T = ten(s); O = ones; m = mental; w = written.

*In the test manual the task structure is described as a + ? = H, we here chose “H00” instead of “H” to make clear that “full” hundreds are meant, like in Item 23.

Table 3: Task Structure of the LVD-M 2-4 Booklets, Grade 4

Item(s)	Task structure	Place value structure/arithmetic operator	Mode	Example
1	a + b = ?	ThHTO + TO (with ? < 10.000)	m	2557 + 43 = ?
2, 3	c – b = ?	10.000 – ThH > 0; ThHTO – ThHT > 0	m	10000 – 6600 = ? 6728 – 4670 = ?
4	a + ? = c	ThHTO + HT (with ? < 10.000)	m	8243 + ? = 9023
5	a + ? = 10.000	ThHTO + (Th)HTO = 10.000	m	5862 + ? = 10000
6	c – ? = a	ThHTO – ThH > 0	m	7536 – ? = 4936
7, 8	a + b = ?	ThHTO* + ThHTO (with carry over once); TThThHTO* + TThThHTO = X0 000 (with X < 10)	w	4157 + 2839 = ? 30818 + 19182 = ?
9	a + b + c = ?	ThHTO** + ThHTO + ThHTO (with carry over once or twice)	w	1092 + 3261 + 2516 =?
10–12	c – b = ?	ThHTO** – ThHTO (with no carry over); TThThHTO** – TThThHTO (with carry over once or twice); 1 000 000 – HThTThThHTO* > 0	w	6898 – 5267 = ?, ..., 1000000 – 403182 = ?

13–18	a · b = ?	To · O;
		TO · O;
		HTO* · O;
		TO · T;
		HTO* · T;
		HTO · TO
			w	30 · 9 = ?, ...,
				275 · 60 = ?

19–24	c : b = ?	HTO : O;
		ThHTO : O;
		TThThHTO* : O;
		ThHT* : To;
		TThTHT : T;
		HThTThThHTO* : T
			w	355 : 5 = ?, ...,
				404580 : 60 = ?

Note. HTh = hundred thousand(s); TTh = ten thousand(s); Th = thousand(s); H = hundred(s); T = ten(s); O = ones; m = mental; w = written.

*One (but the first) place may be zero. **One or two (but the first) place(s) may be zero.

Individual booklets are generated by means of stratified item sampling, that is, each booklet follows the same task and place value structure as well as order of the 24 item families. The specific representation within each of the item families is randomly drawn, assisted by software. This means that the term “item family” refers to the structural pattern of a task (as given in the lines of Table 2). For example, Item Family 3 in Grade 3 booklets (c – b = ?, HTO – O (with ? > H, first O < second O)), could likewise be represented by the random representations “432 – 6 = ?” or “824 – 5 = ?”.

LVD-M 2-4 is designed as a power test with no specific time limit. Administration in classroom settings takes approximately 15 to 20 minutes. Based on the norm sample, the test authors report split-half reliabilities of .87 (Grade 3) and .79 (Grade 4) for measurements halfway through and .81 (Grade 3) and .83 (Grade 4) at the end of the school year (Strathmann & Klauer, 2012). In terms of criterion validity, LVD-M 2-4 scores substantially correlate with results on an established curriculum-based test (DEMAT; correlations between .53 in Grade 3 and .80 in Grade 4) and math grades (correlations between –.54 and –.77; Strathmann & Klauer, 2012, p. 32).

Factor analyses of the norm sample data show that, when the 24 items are transformed to four subscores, one for each basic arithmetic operation, a strong general factor explains between 51% and 78% of the variance in Grade 3 and 4 results (Strathmann & Klauer, 2012). Therefore, to evaluate the test, the number of correctly solved items per student across the whole booklet are counted. This means that the test score x (a) aggregates the different tasks to a general “competence to perform basic arithmetic operations” (Strathmann & Klauer, 2012, p. 31) and (b) does not distinguish between incorrectly solved and unsolved items (both are scored with 0). Based on the binomial model, the ability p of a student to deal with the content represented by the test is estimated as $p = x / n$, with n denoting the number of items and x the number of correctly solved items. While n is a constant, x is a binomially distributed random variable with the variance $s^2 = n \cdot p \cdot (1 - p)$, so students with either very low or very high ps have more precise ability estimates.

2.3 Design

Starting in January and ending in early June 2016 (second half-year), LVD-M 2-4 was conducted bi-weekly, adding up to ten measurements. In all participating classes, the first assessment in January was led by a student assistant of the research team who advised each math teacher to administer the consecutive nine assessments. Teachers received an information brochure with background information on LVD-M 2-4, the schedule, as well as a detailed, standardized instruction sheet. Every other week, teachers were provided with the set of tests to be completed by their class in the following assessment week (the specific assessment day was up to teachers) along with a reply envelope.

All students went through ten different booklets. The booklets (A–H) were presented in ten different orders (1–10) based on a Latin square design (Table 4), and each student was randomly assigned one of the presentation orders. To ensure that data strictly followed this design, booklets were labelled with students' names and teachers were asked to return them directly after completion. Students participated on average on 9.35 of ten measurements (*SD* = 0.96), 54.63% of the participants in Grade 3 and 57.8% in Grade 4 completed all ten booklets.

After five measurements, that is, in spring 2016, all teachers were provided with intermediate feedback on (a) their respective class in comparison to other classes participating in the project as well as for (b) each student compared to the specific classroom. For both levels, class and students, the development was rated as *positive trend*, *negative trend*, or *constant*. In order to classify an individual student trajectory, at least four data points had to be present, and Cohen's f^2 in a linear regression model (dependent variable: LVD-M 2-4 score, independent variable: measurement time) had to correspond at least to a small effect size (Cohen's $f^2 \geq .02$).

Table 4: Measurement Scheme

	Measurements 1–10 / Booklet versions A–J										
	1	2	3	4	5	Easter break	6	7	8	9	10
Order	Jan	Jan	Feb	Feb	Mar		Apr	Apr	May	May	Jun
1	A	B	C	D	E	Feedback	F	G	H	I	J
2	B	C	D	E	F		G	H	I	J	A
3	C	D	E	F	G		H	I	J	A	B
…											
9	I	J	A	B	C		D	E	F	G	H
10	J	A	B	C	D		E	F	G	H	I

2.4 Booklet Effects

2.4.1 Booklet Effects on the Total Score

The total score y_{ijkt} (number of correctly solved items out of 24) of child i in class j at time t completing booklet k was modeled by a fixed intercept and a fixed slope estimating the linear development across the ten bi-weekly measurements, as given in Model (1). This model and its extensions are practically relevant as the total score is the parameter used for test evaluation and interpretation.

Random effects were included to represent the following variance components: First, random intercepts on the level of (a) booklets (b_{0k}), (b) classes (b_{0j}), and (c) individuals ($b_{0i|j}$) were specified to capture differences in the (baseline) test score due to each of the three levels. Second, the variances of random slopes on the class (b_{1j}) and individual level ($b_{1i|j}$) capture differences in linear growth across time between the instances of each level. Correlations between random intercepts and slopes represent the relation between (baseline) level of achievement and development across time within classrooms or individuals. Based on restricted maximum likelihood estimation (REML), Model (1) was fitted separately for each of the two grades.

$$y_{ijkt} = \beta_0 + b_{0k} + b_{0j} + b_{0i|j} + (\beta_1 + b_{1j} + b_{1i|j}) \times time_t + s_{ijkt} \qquad (1)$$

2.4.2 Item (Family) Effects on Solution Probability

By means of four successive models, the equivalence of randomly sampled booklets was not only assessed for the total score as in (1) but also on the item level. All four models are hierarchical logistic regressions (i.e., a multilevel generalized mixed model with logit link), with the logit of a correct solution denoted as $\eta = log\left(\frac{p(Y=1)}{1-p(Y=1)}\right)$.

First, a baseline Model (2) ignoring possible booklet effects served as a benchmark. This model includes the average item difficulty (fixed intercept β_0), the fixed slope, as well as the random intercept (b_{0i}) and the random slope per child i across measurements (b_{1i}). Pairs of random effects (b_{0i}, b_{1i}) are assumed to follow a bivariate normal distribution, that is, their correlations are not restricted. This model is conceptually closest to the binomial model.

$$\eta_{it} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) \times time_{\mathrm{t}} \qquad (2)$$

Second, in a consecutive Model (3), the outcome η_{itc} was predicted by a fixed (β_1) and child-specific random slope (b_{1i}) representing the students’ development over the ten measurements points. The fixed effect δ_c, with c indicating the number of the respective family, expresses the average difficulty of each of the 24 item families and replaces the fixed intercept β_0 of Model (2).

$$\eta_{itc} = \delta_c + b_{0i} + (\beta_1 + b_{1i}) \times time_t \tag{3}$$

Third, to examine the variability of item family effects, random intercepts b_{0l} for the 240 different items l (240 = 24 item families times 10 booklets) were added to Model (3). In this Model (4), the fixed effect $\delta_{c(l)}$ treats all ten randomly drawn members of a specific item family as if they were different items. The item-level random intercept b_{0l} was assumed to be independent of all other random effects and to be N(0,ω_2)-distributed. In Model (4), the variance parameter ω_2 expresses a "global variance" which is independent of the item families.

$$\eta_{ilt} = \delta_{c(l)} + b_{0l} + b_{0i} + (\beta_1 + b_{1i}) \times time_t \tag{4}$$

Fourth, to investigate even more fine-grained item effects, Model (4) was extended by a random intercept for each item clustered by booklets ($b_{0l,c(l)}$, again normally distributed). In other words, Model (5) assesses to what extent items within one family (task type) are homogeneous or heterogeneous across the ten randomly generated booklets. Thus, in Model (5), the item variance depends on the item families.

$$\eta_{ilt} = \delta_{c(l)} + b_{0i} + b_{0l,c(l)} + (\beta_1 + b_{1i}) \times time_t \tag{5}$$

3. Results

3.1 Booklet Invariance in a Sample Receiving Regular Instruction

3.1.1 Booklet Effects on the Total Score

In Grade 3 (n = 108), the variance of random intercepts (level effects) between children (var[$b_{0i|j}$] = 25.722, Table 5) was over 100 times larger than the variance between booklets (k = 10, var[b_{0k}] = 0.245, Table 5). Moreover, also the class-level effect (var[b_{0j}] = 0.163, Table 5) measured only a small proportion (0.6%) of the variance on the individual level. On the class level, random intercepts (b_{0j}) and slopes (b_{1j}) correlated positively (r = .45), meaning that higher class-level baseline scores tended to go along with steeper slopes (so-called Matthew effect). On the individual level, the correlation between random intercepts ($b_{0i|j}$) and slopes ($b_{1i|j}$) was negative, that is, individuals with a higher (baseline) achievement level showed a flatter progress (r = –.61).

In Grade 4 (n = 109), the effects reported for Grade 3 replicated in direction but varied in size: The variance of random intercepts between individuals was almost 50 times larger than between booklets. While the booklet-level variance of random intercepts was similar in both grades (var[b_{ok}] = 0.25 in Grade 3 vs. 0.28 in Grade 4), individual-level variance was lower in Grade 4 (var[$b_{oi|j}$] = 13.80 vs. 25.72), meaning that individual baseline levels were more heterogeneous in Grade 3. Again, the class-level variance of random intercepts only measured a small proportion (4.2%) of the individual-level variance. The class-level Matthew effect was more pronounced in the Grade 4 sample based on mainly written arithmetic booklets (r = .55) compared to Grade 3 where booklets contain mainly mental arithmetic tasks (r = .45, see above). These effects went along with a lower mean baseline (fixed intercept) in Grade 4 (9.31) compared to Grade 3 (15.39). On the individual level, the correlation between random intercepts and slopes was negative but smaller (r = –.43).

Table 5: Variances and Correlations of Random Effects in Model (1)

Level	Parameter	Grade 3 (n = 108)	Grade 4 (n = 109)
	Fixed effects		
	β_0	15.38 (0.54)	9.31 (0.47)
	β_1	0.18 (0.15)	0.27 (0.12)
	Random effects		
Booklet	var(b_{ok})	0.25	0.28
Class	var(b_{oj})	0.16	0.59
Class	var(b_{1j})	0.20	0.11
Class	cor(b_{oj}, b_{1j})	.45	.55
Individual	var($b_{oi|j}$)	25.72	13.80
Individual	var($b_{1i|j}$)	0.22	0.23
Individual	cor($b_{oi|j}$, $b_{1i|j}$)	–.61	–.43
Individual	var(Residual)	18.67	9.41

Note. The standard error of estimation (fixed effects) is indicated in parentheses.

3.1.2 Item (Family) Effects on Solution Probability

Table 6 displays an overview of the four models expressing item family effects: In both grades, Model (3) containing an item-family-specific fixed effect δ_c showed a better fit than the baseline Model (2), indicating that difficulty estimates vary visibly between item families (Grade 3: $\chi^2[23]$ = 2555.8, p < .001, Grade 4: $\chi^2[23]$ = 2743.4, p < .001). In Grade 3, Item 9 (times 9, cf. Table 2) was the easiest (δ_c = 2.79, Table 7) while Item 24 (written subtraction: 1000 – HTO) was the hardest (δ_c = –0.96). Transformed to the more intuitive level of solution probability this means, for example that halfway through the measurement period (measurement $\text{time}_t = 5$) in Grade 3 items with δ_c below –0.65 had a probability of correct solution of less than 50%, items with δ_c above –.65 a solution probability of more than 50%, and items with $\delta_c > .23$ a solution probability of over 95%. In Grade 4, Item 13 (T · O) was the easiest ($\delta_c = 2.27$, Table 7) and Item 24 (HThTThThHTO* : T) the hardest ($\delta_c = -1.81$, Table 7).

The main result contributed by Model (4) is that the variance of the random intercepts clustered by item families, that is, task types, was only one eighth (Grade 3: $\text{var}[b_{0l}] = 0.09$, Table 8) or one tenth (Grade 4: $\text{var}[b_{0l}] = 0.12$, Table 8) of the variance related to individual subjects (Grade 3: $\text{var}[b_{0i}] = 0.72$, Grade 4: $\text{var}[b_{0i}] = 1.18$, Table 8). Although the additional variance component significantly improved model fit (Grade 3: $\chi^2[1] = 61.30$, $p < .001$, Grade 4: $\chi^2[23] = 137.49$, $p <.001$), the effect was relatively small (cf. AIC in Table 6). This means that general level differences in achievement are far more heterogeneous between students than between items.

In both grades, difficulty effects (random intercepts) of the 24 items representing 24 tasks differed in the extent to which they varied across booklets (comparison of Models [4] and [5]: Grade 3: $\chi^2[23] = 46.26$, $p < .01$; Grade 4: $\chi^2[23] = 70.47$, $p < .001$). While some items produced homogeneous difficulties, that is, variances estimated close to zero, across the ten booklets (Grade 3: Items 3, 8, 13, 14, 21, 23, and 24; Grade 4: Items 5, 9, 11, 19, 20, 21, and 22; Table 9), others varied more strongly (Grade 3: e.g., Items 7, 9, 10, 11, 17, and 22; Grade 4: Items 1, 3, 4, 6; Table 9). In Grade 3, the most homogeneous items comprised all four arithmetic operations, four of them were mental, three written tasks. The most heterogeneous items were mainly mental and multiplication or division tasks. In Grade 4, the most homogeneous items were all but one (Item 5) written tasks, requiring addition, subtraction, and division. The four most heterogeneous ones were mental tasks. While in Grade 4 the most difficulty-heterogeneous items could be found in the initial part of the booklet (first quarter), in Grade 3 difficulty-variant items were spread across the entire booklet.

The correlation between average difficulty and cross-booklet heterogeneity (in difficulty) of items, $r(\delta_{c(l)}, \text{var}[b_{0l,c(l)}]$, was $r = .29$ ($p = .16$) in Grade 3 and $r = .06$ ($p = .77$) in Grade 4.

Table 6: Comparison of Models Expressing Item (Family) Effects on Solution Probability

Model	Formula	Description	AIC (*p* model comparison)	
			Grade 3 (n = 108)	Grade 4 (n = 109)
(2)	$\eta_{it} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) \times time_t$	Baseline model ignoring possible booklet effects (benchmark).	18 208	20 904
(3)	$\eta_{itc} = \delta_c + b_{0i} + (\beta_1 + b_{1i}) \times time_t$	The fixed effect δ_c expresses the average difficulty of each of the 24 item families.	15 698 (< .001)	18 206 (< .001)
(4)	$\eta_{ilt} = \delta_{c(l)} + b_{0l} + b_{0i} + (\beta_1 + b_{1i}) \times time_t$	Random intercepts b_{0l} for the 240 different items l (240 = 24 item families times 10 booklets) express item family effects.	15 638 (< .001)	18 071 (< .001)
(5)	$\eta_{ilt} = \delta_{c(l)} + b_{0i} + b_{0l,c(l)} + (\beta_1 + b_{1i}) \times time_t$	The random intercept $b_{0l,c(l)}$ for each item clustered by booklets assesses to what extent items within one family (task type) are homogeneous or heterogeneous across the ten randomly generated booklets.	15 638 (< .01)	18 047 (< .001)

Note. Meaning of the model parameters in the order of appearance: η = logit of a correct solution (link level); β_0 = average item difficulty (fixed intercept); b_{0i} = child-specific random intercept; b_{1i} = child-specific random slope; β_1 = fixed slope, $time_t$ = measurement time (1–10); δ_c = item-specific fixed effect (average difficulty of each of the 24 item families); $\delta_{c(l)}$ = item-specific fixed effect (all 10 randomly drawn representations of a family are treated as different items); b_{0l} = item-specific random intercepts (for all 240 different items); $b_{0l,c(l)}$ = random intercept for each item clustered by booklets indices; i = individual; t = time; c = item family; l = booklet; the value in parentheses is the p value of the model comparison (χ^2 difference test) of the given model to the model before, for example of Model (3) compared to Model (2).

Table 7: Fixed and Random Effects of Model (3), Logit Link

Effect (level)	Parameter	Grade 3 (n = 108)	Grade 4 (n = 109)
Fixed slope	β_1	0.13 (0.01)	0.08 (0.01)
Fixed (item)	δ_{Item1}	1.73 (0.14)	0.95 (0.14)
Fixed (item)	δ_{Item2}	1.96 (0.14)	0.08 (0.13)
Fixed (item)	δ_{Item3}	1.24 (0.13)	−0.25 (0.13)
Fixed (item)	δ_{Item4}	0.59 (0.12)	−0.91 (0.14)
Fixed (item)	δ_{Item5}	0.95 (0.13)	−0.42 (0.13)
Fixed (item)	δ_{Item6}	0.95 (0.13)	0.18 (0.14)
Fixed (item)	δ_{Item7}	0.31 (0.13)	1.97 (0.15)
Fixed (item)	δ_{Item8}	2.71 (0.18)	1.85 (0.15)
Fixed (item)	δ_{Item9}	2.79 (0.19)	1.52 (0.14)
Fixed (item)	δ_{Item10}	2.20 (0.16)	1.72 (0.15)
Fixed (item)	δ_{Item11}	1.78 (0.15)	−0.08 (0.13)
Fixed (item)	δ_{Item12}	1.84 (0.15)	−0.56 (0.13)
Fixed (item)	δ_{Item13}	1.16 (0.13)	2.27 (0.16)
Fixed (item)	δ_{Item14}	2.25 (0.16)	0.65 (0.13)
Fixed (item)	δ_{Item15}	2.11 (0.16)	0.63 (0.14)
Fixed (item)	δ_{Item16}	1.81 (0.15)	0.25 (0.13)
Fixed (item)	δ_{Item17}	2.25 (0.16)	−0.11 (0.13)
Fixed (item)	δ_{Item18}	1.21 (0.14)	−0.91 (0.14)
Fixed (item)	δ_{Item19}	1.30 (0.13)	0.31 (0.15)
Fixed (item)	δ_{Item20}	0.17 (0.12)	−0.14 (0.15)
Fixed (item)	δ_{Item21}	1.12 (0.13)	−0.24 (0.16)
Fixed (item)	δ_{Item22}	0.05 (0.12)	−1.39 (0.18)
Fixed (item)	δ_{Item23}	-0.96 (0.12)	−1.74 (0.21)
Fixed (item)	δ_{Item24}	-0.96 (0.12)	−1.81 (0.23)
Random intercept (individual)	var(b_{0i})	0.69	1.14
Random slope (individual)	var(b_{1i})	0.01	0.01
Correlation of random intercept and slope (individual)	cor(b_{0i}, b_{1i})	.22	−.23

Note. The standard error of estimation (fixed effects) is indicated in parentheses; the structure of Items 1–24 is detailed in Tables 2 and 3.

Table 8: Random Effects of Model (4)

Level	Parameter	Grade 3 (n = 108)	Grade 4 (n = 109)
Item	var(b_{0l})	0.094	0.122
Individual	var(b_{0i})	0.723	1.176
Individual	var(b_{1i})	0.008	0.005
Individual	cor(b_{0i}, b_{1i})	.21	−.22

Table 9: Variances of Model (5)

	Variance of the random intercept for each item clustered by booklets: var($b_{0l,c(l)}$)	
Item family c	Grade 3	Grade 4
1	0.16	0.41
2	0.06	0.04
3	0.00	0.46
4	0.18	0.21
5	0.06	0.00
6	0.02	0.84
7	0.25	0.03
8	0.00	0.03
9	0.37	0.00
10	0.43	0.12
11	0.23	0.00
12	0.19	0.01
13	0.00	0.14
14	0.00	0.07
15	0.13	0.09
16	0.09	0.11
17	0.25	0.08
18	0.01	0.11
19	0.04	0.00
20	0.08	0.00
21	0.00	0.00
22	0.31	0.00
23	0.00	0.01
24	0.00	0.06
Child-level	Grade 3	Grade 4
b_{0i}	0.73	1.18

Note. The structure of Items 1–24 is detailed in Tables 2 and 3.

3.2 Replication Studies in Two Intervention Groups

So far, this paper concentrated on a group of 108 third and 109 fourth graders who received regular instruction during a 19 weeks progress-monitoring phase. This waiting control group of a larger RTI evaluation study represents classroom practice without systematic intervention.

To test and discuss the generalizability of the findings on booklet equivalence, the same analyses were also carried out with the two intervention groups of the same study. In the first of these groups (Intervention Group 1), children with arithmetic performance at risk level or below (PR ≤ 25) received individual computer-based training with a dyscalculia training app throughout the 19 weeks intervention phase. In the second group (Intervention Group 2), a three-tiered scheme was realized: Children in the risk range (25 ≤ PR ≤ 10) worked with the training app, while children with arithmetic performance below PR 10 received the training app and an additional weekly small group in-person training. Taken together, the two intervention groups represent a population of children receiving needs-based intervention beyond regular classroom instruction: In both intervention groups and grades, the ratio of between-children variance (var[$b_{0i|j}$]) and between-booklets variance (var[b_{0k}]) was even more pronounced than in the waiting control group: In Intervention Group 1 (training app), the between-children variance was 229 times (Grade 3) or 252 times (Grade 4) larger than the between-booklet variance (Table A1). In Intervention Group 2 (training app + small group intervention), this ratio was 362 (Grade 3) and 53 (Grade 4), respectively (Table A1).

Beyond booklet equivalence, results of the intervention and waiting control groups differed in the following ways: First, class-level variance of the random intercepts (var[b_{0j}]) was larger in the intervention groups compared to the waiting control group (Table A1), pointing to more heterogeneity between classes in the intervention samples. Second, the correlational patterns of random effects were different in the intervention groups: While random intercepts and slopes on class level were positively correlated in the waiting control group (cor[b_{0j}, b_{1j}] = .45 and .55, Table 5), the opposite was the case for Grade 3 students in both intervention groups (cor[b_{0j}, b_{1j}] = −.75 and −.74, Table A1), and Grade 4 students in Intervention Group 2 (cor[b_{0j}, b_{1j}] = −.57, Table A1). Item-level variances were even smaller than in the waiting control group, with a maximum of 0.08 (Item 16, Grade 3) and 0.05 (Item 10, Grade 4) in the first intervention group and 0.20 (Item 9, Grade 3) and 0.08 (Item 22, Grade 4) in the second intervention group (Table A3).

4. Discussion

4.1 Summary and Relevance

This study is based on an established German math CBM instrument which was applied in a third- and fourth-grade sample receiving regular instruction during a school half-year with ten bi-weekly measurements. The instrument, LVD-M 2-4 (Strathmann & Klauer, 2012), is broadly relevant, since it is easily applicable in classroom contexts and valid for the whole achievement range within the German elementary school curriculum.

Against this background, the present study tested booklet effects to examine to what extent the theoretically justified item-sampling approach of LVD-M 2-4 is empirically backed up. Analyses with linear mixed models showed that in both grades the variance of test scores attributable to different booklets was so small in relation to the between-student variance to be practically irrelevant. This is an important prerequisite for the practical interpretation of repeatedly measured (longitudinal) test scores valid measures of individual learning. Generalized linear mixed models revealed that item difficulty varied within some of the item families, while others showed close to zero variance across the ten booklets. Although, on a descriptive level, heterogeneity tended to correlate with item difficulty in Grade 3 and was more pronounced for mental compared to written tasks in Grade 4, no clear pattern explaining this source of variance emerged. Importantly, the effect was insubstantial in comparison with the effects due to individual differences between students. To learn more about the inconclusive pattern of differences in across-booklet heterogeneity (of difficulties) for the different item families (Table 9), quasi-experimental studies could help to unravel content (task type, arithmetic operation, etc.) from sequence effects (position, order).

Taken together, it seems legitimate to assume that the t = 10 measurements of LVD-M 2-4 overall express the same ability. This finding is relevant for the *conceptualization perspective* on progress-monitoring instruments and the first stage of CBM research described by Fuchs (2004): Like in previous research (Christ & Vining, 2006), it corroborates that stratified random item-sampling within predefined competences (e.g., written or mental arithmetic, with a certain task and place value structure; cf. Table 2 and Table 3) leads to comparable parallel tests, which is a key element of generic test construction (Rohwer, 2015).

The two dyscalculia intervention groups of the larger RTI study differed from the regular-instruction group in some respect, for example, showing a reversed Matthew effect on the class level (weaker start level along with higher improvement), which might be to some degree explainable by a compensation effect caused by the intervention framework. However, the main finding of this study, that is, a relatively low proportion of test score variance accounted for by different booklets compared to different individuals, clearly generalized to two intervention samples. In other words, equivalence of LVD-M 2-4 booklets can also be assumed for RTI settings with in-depth fostering of students with special educational needs in basic arithmetic.

This is an important precondition for further research aims, for example, assessing sensitivity to change or deriving norms for the learning slope. Such follow-up goals are only reasonable if they build on solid grounds in terms of booklet equivalence.

4.2 Limitations and Further Directions

The results of this study (negligible booklet effects) are limited by its context: the methodological approach, the instrument, and possibly also the population underlying the regional sample in this study.

Regarding the methodological approach, in the current study, we did not proceed by calibrating all items a priori and then investigating DIF or measurement invariance. Rather, test booklets were randomly allocated to students, corresponding to the random groups design as outlined by Kolen and Brennan (2014). According to this approach, "If the same scaling convention […] for ability is used […], then the parameter estimates […] are assumed to be on the same scale without further transformation" (p. 182). Hence, we assumed that all test scores and item parameters were on a comparable scale.

We utilized a statistical approach similar to generalizability theory (Brennan, 2001) in which we compared the magnitude of variance components (cf. Fan & Hansmann, 2015) to establish empirical evidence of test equivalence. This approach does not focus on the equivalence of single item parameters (e.g., classical DIF or measurement invariance approaches). Instead, the approach chosen here deals with the variance of relevant parameters or factors (e.g., item parameters, test scores). In the case of test equivalence, the variance component pertaining to booklets or items should be very small and negligible, compared to factors that should substantially contribute to total variance (e.g., differences in person ability, time points). In fact, the complete absence of item (family)-level and booklet-level variance would imply that parameters of items in the same position are stable between booklets. The tests would then have a strong form of measurement equivalence, in the sense that there was neither DIF nor DTF and the test would parallel in the sense of CTT.

Although the results provided here lend support to the assumption that test forms were practically equivalent, we recognize the limitations of the chosen statistical approach. Specifically, by focusing on variance components, we did not look at single item parameters, which would be helpful in identifying those items whose presence substantially enlarges the family-wise variance component. Further, it remains difficult to provide a suitable effect size of test or item equivalence in this context. At what (relative) test form variance component magnitude can test equivalence be assumed, and when does it not hold? Does test form equivalence obtained using an approach comparing variance components also hold under a DIF or measurement invariance approach? These questions necessitate further research.

Regarding the instrument (LVD-M 2-4) and its construction rationale, we investigated the equivalence of math progress-monitoring tests that were constructed using a systematic rule-based item design approach. In rule-based item design, item features that should theoretically affect item difficulty are identified (Enright et al., 2002). Once these essential item features are known, design of equivalent test forms becomes more straightforward.

However, available progress-monitoring tests differ in how thoroughly essential item features were identified and/or taken into consideration in test design. Some progress-monitoring measures are based on comprehensive and detailed, rule-based item design (e.g., Förster & Kuhn, 2021; Klauer & Strathmann, 2012), providing evidence for a high degree of test equivalence. Other progress-monitoring measures are less explicit concerning the rationale for test and item design. The results obtained in this study, therefore, should not be seen as representative for progress-monitoring measures in general. Rather, they should be regarded as preliminary evidence of equivalence of a progress-monitoring measure that was constructed using rule-based item design.

Moreover, to what extent the findings can be generalized to broader content domains (like in *quop*, which includes precurricular basic numerical abilities in early-grade version and aspects like geometry or basic statistics in later grades; Souvignier, 2018) or different administration modes (computer-based instead of paper-pencil) remain empirical questions for subsequent studies.

Acknowledgements

This research was funded by the German Federal Ministry of Education and Research (grant 01 GJ 1302).

References

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. https://doi.org/10.1080/02796015.2009.12087837

Ardoin, S. P., Christ, T. J., Morena, L. S., Cormier, D. C., & Klingbeil, D. A. (2013). A systematic review and summarization of the recommendations and research surrounding Curriculum-Based Measurement of oral reading fluency (CBM-R) decision rules. *Journal of School Psychology, 51*(1), 1–18. https://doi.org/10.1016/j.jsp.2012.09.004

Ashlock, R. B. (2005). *Error patterns in computation: Using error patterns to improve instruction* (9th ed.). Prentice Hall

Björn, P. M., Aro, M., Koponen, T., Fuchs, L. S., & Fuchs, D. (2018). Response-to-intervention in Finland and the United States: Mathematics learning support as an example. *Frontiers in Psychology, 9*, Article 800. https://doi.org/10.3389/fpsyg.2018.00800

Brennan, R. L. (2001). *Generalizability theory*. Springer. https://doi.org/10.1007/978-1-4757-3456-0

Christ, T. J., & Vining, O. (2006). Curriculum-based measurement procedures to develop multiple-skill mathematics computation probes: Evaluation of random and stratified stimulus-set arrangements. *School Psychology Review, 35*(3), 387–400. https://doi.org/10.1080/02796015.2006.12087974

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. (2003). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention, 28*(3–4), 3–12. https://doi.org/10.1177/073724770302800302

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education, 15*(1), 49–74. https://doi.org/10.1207/S15324818AME1501_04

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

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics. *The Journal of Special Education, 41*(2), 121–139. https://doi.org/10.1177/00224669070410020101

Förster, N., & Kuhn, J.-T. (2021). Ice is hot and water is dry – Developing equivalent reading tests using rule-based item design. *European Journal of Psychological Assessment*. Advance online publication. https://doi.org/10.1027/1015-5759/a000691

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*(3), 315–342. https://doi.org/10.1016/j.jsp.2007.06.003

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188–192. https://doi.org/10.1080/02796015.2004.12086241

Gerster, H.-D. (2012). *Schülerfehler bei schriftlichen Rechenverfahren: Diagnose und Therapie* [Student errors in written arithmetic procedures: Diagnosis and therapy]. WTM. https://nbn-resolving.de/urn:nbn:de:hbz:6-79209639973

Hebbecker, K., & Souvignier, E. (2016). Lernverlaufsdiagnostik zur Unterstützung individueller (Begabungs-)Förderung: Internetbasierte Lernverlaufsdiagnostik mit dem System „quop" [Learning progress diagnostics to support individual (gifted) support: Internet-based learning progress diagnostics with the system "quop"]. *Journal für Begabtenförderung, 16*(1), 29–38.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABC's of CBM: A practical guide to curriculum-based measurement* (2nd ed.). Guilford.

Huber, C., & Grosche, M. (2012). Das Response-to-Intervention-Modell als Grundlage für einen inklusiven Paradigmenwechsel in der Sonderpädagogik [The response to intervention model as a basis for an inclusive paradigm shift in special education]. *Zeitschrift für Heilpädagogik, 63*(8), 312–322.

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review, 37*(3), 374–390. https://doi.org/10.1080/02796015.2008.12087884

Klauer, K. J. (2011). Lernverlaufsdiagnostik – Konzept, Schwierigkeiten und Möglichkeiten [Diagnosing the course of learning – Concept, difficulties and chances]. *Empirische Sonderpädagogik, 3*(3), 207–224.

Klauer, K. J. (2014). Formative Leistungsdiagnostik: Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik [Formative achievement assessment: Historical background and evolution into progress monitoring]. In M. Hasselhorn, W. Schneider, & U. Trautwein (Eds.), *Tests und Trends: Vol. 12: Lernverlaufsdiagnostik* (pp. 1–18). Hogrefe.

Klauer, K. J., & Strathmann, A. M. (2013). Lernverlaufsdiagnostik Mathematik: Test auf Änderungssensibilität bei rechenschwachen Grundschülern [Assessing growth in mathematical achievement: A test of sensitivity to change with underperforming elementary students]. *Psychologie in Erziehung und Unterricht, 60*(4), 241–252. https://doi.org/10.2378/peu2013.art18d

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7

Köller, O., & Reiss, K. (2013). Mathematische Kompetenz messen: Gibt es Unterschiede zwischen standardbasierten Verfahren und diagnostischen Tests? [Assessing mathematics competence: Are there differences between standard-based methods and diagnostic tests?] In M. Hasselhorn, A. Heinze, W. Schneider, & U. Trautwein (Eds.), *Diagnostik mathematischer Kompetenzen* (pp. 25–40). Hogrefe.

Krajewski, K., Liehm, S., & Schneider, W. (2004). *Deutscher Mathematiktest für zweite Klassen (DEMAT 2+)* [German mathematics test for second grades]. Hogrefe.

Kuhn, J.-T., & Holling, H. (2014). Number sense or working memory? The effect of two computer-based trainings on mathematical skills in elementary school. *Advances in Cognitive Psychology, 10*(2), 59–67. https://doi.org/10.5709/acp-0157-2

Kuhn, J.-T., & Schwenk, C. (2018). Onlinebasierte Diagnostik mathematischer Kompetenzen: Möglichkeiten und Grenzen [Online-based assessment of mathematical competencies: Possibilities and limitations]. *Lernen und Lernstörungen, 7*(4), 231–235. https://doi.org/10.1024/2235-0977/a000232

Lawson, M. J., & Rice, D. N. (1987). Thinking aloud: Analysing students' mathematics performance. *School Psychology International, 8*(4), 233–244. https://doi.org/10.1177/0143034387008004004

McDonald, R. P. (1999). *Test theory: A unified treatment*. Erlbaum.

Montague, M., Penfield, R. D., Enders, C., & Huang, J. (2010). Curriculum-based measurement of math problem solving: A methodology and rationale for establishing equivalence of scores. *Journal of School Psychology, 48*(1), 39–52. https://doi.org/10.1016/j.jsp.2009.08.002

Rohwer, G. (2015). Bemerkungen zu einem Testverfahren für Lernfortschritte [Remarks on a test procedure for long-term learning progress]. *Journal for Educational Research Online, 7*(2), 147–156.

Roick, D., Görlitz, T., & Hasselhorn, M. (2004). *Deutscher Mathematiktest für dritte Klassen (DEMAT 3+)* [German mathematics test for third grades]. Hogrefe.

Schatschneider, C., Wagner, R. K., & Crawford, E. C. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning and Individual Differences, 18*(3), 308–315. https://doi.org/10.1016/j.lindif.2008.04.005

Schwenk, C., Kuhn, J.-T., Gühne, D., Doebler, P., & Holling, H. (2017). Auf Goldmünzenjagd: Psychometrische Kennwerte verschiedener Scoringansätze bei computergestützter Lernverlaufsdiagnostik im Bereich Mathematik [We are going on a gold coin hunt: Psychometric properties of different scorings in computer-based progress monitoring of mathematics ability]. *Empirische Sonderpädagogik, 9*(2), 123–142.

Shapiro, E. S., Dennis, M. S., & Fu, Q. (2015). Comparing computer adaptive and curriculum-based measures of math in progress monitoring. *School Psychology Quarterly, 30*(4), 470–487. https://doi.org/10.1037/spq0000116

Souvignier, E. (2018). Computerbasierte Lernverlaufsdiagnostik [Computer-based learning progress assessment]. *Lernen und Lernstörungen, 7*(4), 219–223. https://doi.org/10.1024/2235-0977/a000240

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*(3), 407–419. https://doi.org/10.1080/02796015.2001.12086123

Strathmann, A. M., & Klauer, K. J. (2010). Lernverlaufsdiagnostik: Ein Ansatz zur längerfristigen Lernfortschrittsmessung [Diagnosing the trajectory of learning: An approach to long term measuring of learning progress]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42*(2), 111–122. https://doi.org/10.1026/0049-8637/a000011

Strathmann, A. M., & Klauer, K. J. (2012). *LVD-M 2-4. Lernverlaufsdiagnostik – Mathematik für zweite bis vierte Klassen* [LVD-M 2-4: Learning-progress diagnostics – mathematics for grades two to four]. Hogrefe.

Voß, S. (2016). Rechengeschwindigkeit, -präzision oder -flüssigkeit? Zur Vorhersage und Förderung der Rechenleistungen von Erstklässlern [Computational speed, accuracy or fluency? The prediction and promotion of computational skills of first-grade students]. *Heilpädagogische Forschung, 42*(1), 13–24.

Weiß, R. H. (2006). *CFT 20-R – Grundintelligenztest Skala 2 mit Wortschatztest (WS) und Zahlenfolgentest (ZF)-Revision* [Culture Fair Intelligence Test, Scale 20-R]. Hogrefe.

Weiß, R. H., & Osterland, J. (2013). *CFT 1-R – Grundintelligenztest Skala 1 – Revision* [Culture Fair Intelligence Test, Scale 1]. Hogrefe.

Wimmer, H., & Mayringer, H. (2014). *Salzburger Lese-Screening für die Schulstufen 2–9* [Salzburg reading screening test for grades 2–9]. Hogrefe.

Appendix

Table A1: Variances and Correlations of Random Effects in Model (1) in the Two Intervention Groups

		Intervention group 1		Intervention group 2	
Level	Parameter	Grade 3 (n = 140)	Grade 4 (n = 118)	Grade 3 (n = 103)	Grade 4 (n = 125)
Fixed effects					
	β_0	14.52 (1.02)	8.44 (0.85)	13.18 (1.48)	8.23 (1.06)
	β_1	0.17 (0.16)	0.36 (0.14)	0.10 (0.21)	0.22 (0.12)
Random effects					
Booklet	var(b_{0k})	0.18	0.08	0.10	0.28
Class	var(b_{0j})	6.44	4.09	17.47	8.66
Class	var(b_{1j})	0.46	0.15	0.40	0.11
Class	cor(b_{0j}, b_{1j})	−.75	−.08	−.74	−.57
Individual	var($b_{0i\|j}$)	42.10	20.96	37.61	14.62
Individual	var($b_{1i\|j}$)	0.27	0.19	0.34	0.28
Individual	cor($b_{0i\|j}$, $b_{1i\|j}$)	−.45	.01	−.59	−.26
Individual	var(Residual)	13.89	8.85	23.26	15.17

Table A2: Random Effects of Model (4) in the Two Intervention Groups

		Intervention Group 1		Intervention Group 2	
Level	Parameter	Grade 3 (n = 129)	Grade 4 (n = 104)	Grade 3 (n = 103)	Grade 4 (n = 124)
Item	var(b_{0l})	0.00	0.00	0.00	0.00
Individual	var(b_{0i})	0.70	0.59	1.86	0.98
Individual	var(b_{1i})	0.01	0.00	0.09	0.00
Individual	cor(b_{0i}, b_{1i})	−.11	.20	−.51	−.10

Table A3: Variances of Model (5) in the Two Intervention Groups

	Variance of the random intercept for each item clustered by booklets: var($b_{0l,c(l)}$)			
	Intervention Group 1		Intervention Group 2	
Item family c	Grade 3	Grade 4	Grade 3	Grade 4
1	0.00	0.00	0.03	0.00
2	0.00	0.00	0.00	0.03
3	0.00	0.03	0.00	0.00
4	0.00	0.00	0.00	0.00
5	0.00	0.03	0.00	0.00
6	0.00	0.00	0.00	0.00
7	0.00	0.01	0.00	0.00
8	0.00	0.00	0.03	0.00
9	0.04	0.00	0.20	0.00
10	0.03	0.05	0.00	0.00
11	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.01
13	0.01	0.00	0.04	0.00
14	0.00	0.00	0.00	0.01
15	0.00	0.00	0.00	0.01
16	0.08	0.03	0.00	0.00
17	0.06	0.00	0.00	0.00
18	0.05	0.02	0.00	0.00
19	0.00	0.00	0.00	0.07
20	0.00	0.04	0.00	0.00
21	0.00	0.03	0.04	0.00
22	0.00	0.00	0.00	0.08
23	0.00	0.00	0.00	0.00
24	0.00	0.00	0.00	0.03
b_{0i}	0.70	0.59	1.87	0.978

Note. The structure of Items 1–24 is described in Table 2 and Table 3.