Web-based progress monitoring in first grade mathematics

Martin Salaschek[a], Elmar Souvignier[a]

[a]University of Münster, Germany

Abstract

The purpose of our research was to examine a web-based tool for mathematics progress monitoring in first grade. The newly developed assessment tool uses several robust indicators and curriculum-based measures forming three competences (Basic Precursors, Advanced Precursors, and Computation) to determine comprehensive early numeracy skills in general education. 373 students completed a total of eight online tests every two or three weeks. Results indicate that delayed alternate-form reliability was adequate ($r_M = .78$). Repeated measures analyses with post hoc comparisons were used to ascertain the sensitivity to assess learning growth. All three competences showed linear growth rates that were significant over time, but only Computation and overall scores produced dependable increases from test to test. Predictive validity was determined using two standardised school achievement tests (end of first grade, end of second grade). Results indicate high predictive validity of the first four online tests ($r_M = .67$, $r_M = .66$ for 6 months and 18 months prediction). Correlations with teacher ratings of their students' skills confirmed this pattern. Results from student and teacher questionnaires indicate that the students were able to conduct the tests independently and that a three-week interval was adequate for regular-education use. Teachers declared to use the progress monitoring results diversely for classroom purposes. We conclude that the use of a web-based assessment setting with diverse measures is beneficial with respect to psychometric properties and feasibility for frequent use in general education.

Keywords: Early numeracy; Mathematics; Progress monitoring; Web-based assessment

1. Introduction

Learning progress assessment aims at providing teachers with information about learning growth, and using diagnostic information for individualised instruction has been shown to result in higher learning gains (Connor, Morrison, & Petrella, 2004; Stecker, Fuchs, & Fuchs, 2005). Especially in first grade, results from Kim, Petscher, Schatschneider, and Foorman (2010) show that the slope of learning is highly predictive for future achievement. However, Stecker et al. note that teachers need assistance in interpreting and
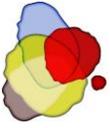
successfully using progress monitoring results. Progress monitoring tools should therefore provide educators with reliable and comprehensive feedback about students' skills. For successful implementation in regular-education classrooms, high utility and feasibility is additionally required. This can be achieved with highly automated assessment and feedback systems. Traditional progress monitoring tools reliably and validly assess students' performance, but are time-consuming because they usually require face-to-face assessment. In addition, most tools for first grade consist of only a few different curricular tasks, making it difficult for educators to use results for adjustments in classroom work. In the present study, we examined psychometric properties and utility of a web-based progress monitoring tool for first-graders. The tool assesses early mathematics competences comprehensively and allows students to work on the tests independently without teacher aid.

1.1 Early numeracy and later mathematical achievement

Early numeracy plays a vital role for the development of later mathematics performance and general school achievement (Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Duncan et al., 2007). Thus, much research in the past decade has focused on the identification of relevant skills that children should be proficient in when entering school (Berch, 2005; Gersten, Jordan, & Flojo, 2005; Jordan, Kaplan, Oláh, & Locuniak, 2006; Koponen, Aunola, Ahonen, & Nurmi, 2007; Methe, Begeny, & Leary, 2011; Missall, Mercer, Martínez, & Casebeer, 2012). Certain *number sense* abilities seem to form precursors or even gateways for further mathematical achievement, but the definition of number sense remains vague (cf. Berch, 2005, for an overview). Unlike reading, in which well-defined precursors (such as phonological awareness) have been identified, numeracy seems to develop from a diverse set of mental processes which evolve during childhood. The *triple-code model of number processing* (Dehaene & Cohen, 1995; Dehaene, 1992, 2011) describes three systems involved in different aspects of number processing (i.e., for nonverbal semantic representations; for verbal representations; and for written numerals) derived from a biological viewpoint. These systems develop independently, and pathways are used for communication when solving mathematical problems. Developmental models like the *model of early mathematical development*, which describes three levels of successional skills (Krajewski & Schneider, 2009; Krajewski, 2008), take up a more growth-oriented stance. In Krajewski's model, skills at the second level represent the linking of number words with quantities. These skills proved to be particularly predictive for mathematical achievement at the end of primary school (Krajewski & Schneider, 2009).

1.2 Progress monitoring in early mathematics

Students at risk of not reaching educational goals can be identified by assessing progress of essential skills, such as curricular abilities and number sense skills, which have been described as "gateway" skills for further mathematical development (Clarke, Baker, Smolkowski, & Chard, 2008, p. 48). Subsequently, suitable interventions can be implemented. Educators can use tools to monitor learning progress over time and thereby identify students who do not improve (at an acceptable rate). Assessment tools for this purpose should reliably assess students' performance level and its development, so that students at risk of not reaching curricular goals can be identified. Furthermore, diagnostic information about curricular competences should be provided, which teachers can use for instructional changes. Implementation should be efficient and as effortless as possible such that general classroom work is not hindered (Förster & Souvignier, 2011).

One progress monitoring approach for this purpose is *Curriculum-Based Measurement* (CBM; see Deno, 2003, for an overview). In CBM, short tests of important curricular competences are conducted regularly. For early mathematics, the psychometric properties of several CBM tests have been discussed in the literature recently (e.g., Chard et al., 2005; Clarke et al., 2011; Seethaler & Fuchs, 2011). Much of the recent early mathematics CBM research focuses on a set of measures known as *Tests of Early Numeracy*
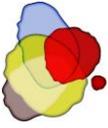
(TEN). TEN measures have demonstrated high levels of reliability and predictive value for later mathematics performance in a number of studies during kindergarten and first grade general education (e.g., Baglici, Codding, & Tryon, 2010; Chard et al., 2005; Clarke & Shinn, 2004; Missall et al., 2012). TEN consist of four measures: (1) *Oral Counting*, assessing the ability to count orally; (2) *Number Identification*, assessing the ability to verbally identify a written number between 0 and 20; (3) *Quantity Discrimination*, assessing the ability to identify the larger of two visually presented numbers; and (4) *Missing Number*, assessing the ability to name the missing number from a string of three numbers, with one of the three numbers missing.

However, there are several issues still to be worked on if these measures shall serve as a basis for instructional changes in the classroom: First, as Methe (2012, p. 68) notes, TEN measures "struggle to capture more exact knowledge deficits" because they lack close relation to curricula. Results are therefore hard to interpret by educators. Measures that relate more closely to specific curricular goals might make it easier for educators to use the diagnostic information for classroom work or further interventions. Second, reliability and predictive validity results of the four single measures vary from study to study (see Missall et al., 2012, for an overview); Missall et al. (l.c., p. 96) ascertain that a combination of several measures seems to result in elevated technical adequacy. As a consequence, the authors call for progress monitoring tools which assess early mathematics more comprehensively. Third, with the recent exception of a study by Hampton et al. (2012), most studies report results from only two or three data points and interpolate learning growth between them. This procedure does not allow a timely evaluation of individual learning growth and also leaves the possibility of non-linear growth patterns. This aspect is especially relevant in the light of low (interpolated) weekly growth rates that often do not exceed 0.30 points per week (Foegen, Jiban, & Deno, 2007). Low average growth rates make it more difficult to interpret stagnating scores as *at-risk*. Finally, TEN measures are time-consuming to implement because two of the measures (Oral Counting and Number Identification) require students to verbalize their answers and therefore can only be assessed in one-on-one settings. In general education, the time and effort needed are reasons why educators usually do not utilise early mathematics progress monitoring at all or regularly enough to make quick instructional adjustments possible.

1.3   Aims of the study

In our study we aim to approach the aforementioned issues with a web-based progress monitoring tool for first grade mathematics which is feasible for frequent use in general education. The tool intends to assess mathematics skills comprehensively and includes both precursor and curricular competences. That way, educators are enabled to make inferences about students' strengths and weaknesses for classroom work or intervention. Assessment time needs to be low and the retrieval and use of results as effortless as possible. Psychometric properties of the test concept should be sufficient for dependable estimations of students' short-term and long-term curricular achievements and for the detection of learning growth. Students should work on the tests in a motivated manner to obtain valid results.

These aims lead to the following research questions: (1) Does the progress monitoring tool assess students' performance reliably? (2) As measures of concurrent and predictive criterion validity, do the progress monitoring test scores correlate significantly with results from standardised achievement tests and teacher ratings of students' mathematics performance? (3) Are learning gains represented in the test scores? I.e., can increases in test scores be observed when testing frequently? (4) Do teachers and students rate the tool and its implementation feasible for frequent use in general education?
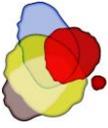
2. Method

2.1 Participants and setting

Two consecutive studies were conducted with a total of 373 first-grade students in 18 regular-education classrooms (see Table 1 for demographics). The studies took place in rural and urban areas of Germany. Eight progress monitoring tests were conducted in both studies in intervals of either two weeks (study 1, November 2010 to March 2011) or three weeks (study 2, November 2011 to May 2012). Figure 1 provides an overview of the time structure and main dependent variables of the two studies.

In study 1, a number of additional measures was obtained: Three different standardised paper-pencil tests (pp1-pp3) were conducted, assessing relevant curricular competences of each time point. pp1 was conducted immediately before the first progress monitoring test, pp2 immediately after the last progress monitoring test. Eight of the 10 classrooms in study 1 (148 students) participated in a follow-up paper-pencil test approximately 14 months later at the end of second grade (pp3). Teacher ratings of students' overall mathematical competence were obtained before each of the three school achievement tests. At the end of first grade, teachers were also surveyed about the feasibility of the web-based progress monitoring tool and their use of the results. Students completed a short questionnaire about the progress monitoring test before pp2.

Purpose of study 1 was to obtain detailed information about the tests' validity. Study 2 was then conducted to inspect reliability and sensitivity to learning in an extended time-frame. In preparation of study 2, single items were revised pertaining to difficulty and parallelism after study 1.

Because of student mobility or sick absentees, some data were missing (progress monitoring tests: 0%-11%, $M_{missing} = 1.8\%$; paper pencil tests: 0%-3.6%, $M_{missing} = 1.7\%$; teacher ratings: 4.5%-23.2%, $M_{missing} = 12.6\%$). We used multiple imputation with five imputed data sets to handle missing test data (Newton et al., 2004). Unbiased results can be expected from multiple imputation when data are missing at random (MAR; see Schafer & Graham, 2002, for a discussion of the term) or when auxiliary variables are included in the imputation model which closely relate to the missing data (Collins, Schafer, & Kam, 2001). Given the number of strongly correlated variables in our study designs, we assumed that our inclusive multiple imputation model produced results that are not meaningfully biased. Where applicable, coefficients reported in the results section were obtained by combining the imputed data sets using the formulas reported by Rubin (1987, 1996).

Table 1

Demographics of study participants

	Study 1	Study 2
N	220	153
Sex		
Girls	51%	46%
Boys	49%	54%
Migration background	22%	9%
Age at first progress monitoring test	6.68 years	6.72 years

Note. Migration background was defined via language(s) spoken at home. Students who spoke another language than German at home were categorized as having a migration background.

	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May		June grade 2
Study 1	pp1	progress monitoring tests 1-8 (2-week intervals)				pp2			pp3
Study 2		progress monitoring tests 1-8 (3-week intervals)							
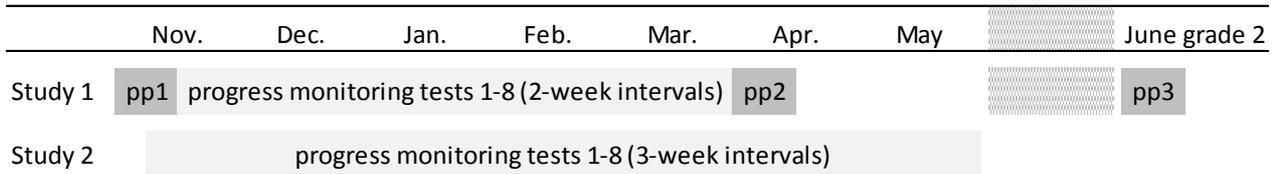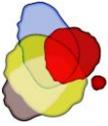
Figure 1. Schematic overview of the time structure of study 1 and study 2. Study 1 was conducted from November 2010 to June 2012, study 2 was conducted from November 2011 to May 2012. pp = paper pencil test.

2.2 Progress monitoring measures

Progress monitoring tests consisted of nine measures in three competences with a total of 52 problems (Table 2 provides an overview of the measures used in the progress monitoring test in both studies). The tests were completely computerised, and students received detailed audio instructions before each new set of tasks via headphones to eliminate the influence of reading skills. All tasks were in multiple choice format, in which students clicked on the solution they thought to be correct. Tests were untimed, and the children worked on them independently without teacher instruction. Results were computed as percentage correct, and educators could access results (graphs and tables) at student and classroom level immediately after a test was completed by a student. Results could be compared with class means or overall mean scores of all participating classrooms in the study, and results differing more than one standard deviation from the mean could be highlighted.

During the two-week/three-week interval of each test, classrooms could choose to test all students during one class period (if computer rooms were available) or consecutively on computers in the classroom, e.g., during self-study periods. A time frame of two weeks per test was initially chosen for particularly close monitoring of learning growth. Intervals were extended to three weeks in study 2 as a response to teacher feedback.

The test emphasized the gateway role of number sense by assessing two sets of precursor skills, *Basic Precursors* and *Advanced Precursors*. Both competences were closely related to the triple-code model (Dehaene & Cohen, 1995) and Krajewski and Schneider's model of early mathematical development (Krajewski & Schneider, 2009). Precursor measures were complemented by relevant curriculum-based *Computation* skills. All measures included questions of varying difficulty to differentiate between weaker and stronger students. Four parallel versions (A-D) of the test were created by using item-cloning algorithms for task creation and the selection of distractors (cf. Clause, Mullins, Nee, Pulakos, & Schmitt, 1998): For every task, attributes that define its difficulty were identified and held constant in the parallel tests (e.g., for an addition task, the size of the second summand and whether crossing the tens boundary was necessary). Throughout the school year, each of the four tests was conducted twice to obtain eight data points (sequence A-D, A-D).

Basic Precursors aimed at assessing fundamental skills that students should be proficient at when entering school. Basic Precursors contained the measures Number Discrimination (similar to the TEN measure Quantity Discrimination), Symbol Quantity Discrimination, and Number Identification (also similar to the corresponding TEN measure).

Advanced Precursors aimed at more sophisticated precursor skills, which usually partly develop before school entrance and should soon be mastered during school. Advanced Precursors contained the measures Number Sequence 1/Number Sequence 2 (similar to the TEN measure Missing Number and the Next Number task used by Hampton et al., 2012) and Number Line, which assesses the extent to which a linear mental number line is developed (see Siegler & Booth, 2004, for a discussion).
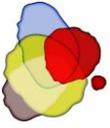
Computation aimed at the main curricular arithmetic goals of German first grade, i.e., handling numbers in the range of 1-20. Computation contained addition and subtraction tasks as well as equation problems with dice.

Table 2

Description of progress monitoring measures

Competence/Measure	No. of items	Range	Example problem	Distractors	Task description
Basic Precursors	*20*				
Number Discrimination	8	1-500	64 | 38		Select the larger number
Symbol Quantity Discrimination	6	1-10			Select the picture with more shapes
Number Identification	6	1-100	*Audio: "28"*	82 | 27 | 72 | 28 | 38	Select the number that was given via audio
Advanced Precursors	*17*				
Number Sequence 1	4	1-20	19, 18, ?	15 | 20 | 16 | 17	Select the missing number (steps of 1)
Number Sequence 2	4	1-20	4, 6, ?	10 | 8 | 9 | 7	Select the missing number (steps of 2)
Number Line	9	1-20	*Audio: "12"*		Select the number line that has a mark at the position of the number that was given via audio
Computation	*15*				
Addition	5	1-20	6 + 5 = ?	9 | 10 | 11 | 13	Select the correct solution
Subtraction	4	1-20	15 - 8 = ?	7 | 9 | 23 | 5	Select the correct solution
Equation	6	1-10		4 + 4 | 7 + 3 | 4 + 3	Select the problem with the same solution as the dice problem
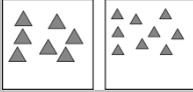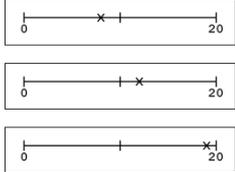
Note. All measures contained problems of varying difficulty, e.g., lower or higher numbers. Detailed task descriptions were provided via headphones in language suitable for children.

2.3 Criterion measures

The three paper-pencil achievement tests in study 1 were selected with reference to their curricular adequacy of the given time points. E.g., at the beginning of grade 1, an achievement test suitable for whole classrooms cannot yet test curricular competences which are only expected to develop during the school year. For this reason, the *Osnabrück test of number concept development* (OTZ; van Luit, van de Rijt, &Hasemann, 2001) was chosen as pp1. The OTZ is suitable for children age 4.5 to 7.5 and assesses
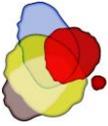
precursor skills such as counting, sorting, and comparing quantities. At the end of first grade, the *German mathematics test for first grade* (DEMAT 1+; Krajewski, Küspert, & Schneider, 2002) was chosen as end-of-year criterion (pp2). The DEMAT 1+ was developed following models of early mathematical development, but mainly assesses curricular goals from first grade, e.g., addition/subtraction in the range of 1-20 and (de)composition of numbers. At the end of second grade, the *German mathematics test for second grade* (DEMAT 2+; Krajewski, Liehm, & Schneider, 2004) was chosen for inspecting long-term predictive validity (pp3). The DEMAT 2+ assesses the main curricular goals from second grade, e.g., basic arithmetic operations in the range of 1-100, number properties, and geometry problems. Paper pencil tests were group-administered within one 45-minute period in all classrooms[1]. All paper-pencil data were collected and put in by trained university students. Results were calculated automatically from raw test answers to prevent scoring errors.

Before each paper pencil test, teachers were asked to rate each of their students' overall mathematic competence on a 7-point Likert scale.

2.4 Usability and practicality

For study 1, several measures of feasibility of the progress monitoring tests were assessed. Students were surveyed about the computer tests after completion of all eight probes, asking (1) how they liked the tests, and (2) how they would like to do more tests in the next school year. A 5-point Likert scale using smiley faces was used as answer format. Additionally, as a measure of direct usability, the time needed to complete each test was logged by the test system. Finally, all 10 teachers from study 1 completed a survey about implementation time and their usage of test results.

3. Results study 1

3.1 Internal reliability

We computed the internal reliability for total scores and the three competences. Mean reliability of total scores was .86 and varied within a narrow range, demonstrating good overall internal consistency. Reliabilities of the single competences were lower: While Advanced Precursors showed satisfactory reliability, coefficients of Basic Precursors and Computation ranged from low to acceptable (see Table 3).

---

[1] OTZ tasks were slightly adjusted to allow group administration (no German standardised paper pencil test that originally allows group administration was available). For DEMAT 1+ and DEMAT 2+, one task was omitted that had not been introduced in any of the participating classes at the time of testing. Thus, overall results are not directly comparable to the reference sample reported by the test authors.
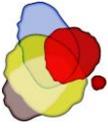
Table 3

Internal consistencies of progress monitoring overall scores and competence scores

progress monitoring	α			
	Overall score	Basic Precursors	Advanced Precursors	Computation
time 1	.84	.65	.72	.71
time 2	.86	.60	.78	.71
time 3	.85	.62	.79	.69
time 4	.85	.55	.81	.74
time 5	.87	.65	.83	.74
time 6	.86	.64	.80	.76
time 7	.88	.66	.82	.79
time 8	.88	.65	.84	.79
M	.86	.63	.80	.74

3.2 Concurrent and predictive validity

3.2.1 School achievement tests[2]

As a measure of concurrent validity, correlations between the progress monitoring tests and grade 1 fall pp1 scores were moderate, with $.40 \leq r \leq .50$. To assess the progress monitoring tests' capacity to predict later mathematics performance early in the school year, correlations between the first four tests and grade 1 spring pp2 scores were calculated. Coefficients were higher, with $.64 \leq r \leq .71$, indicating strong predictive validity for the end-of-year performance. Correlations between the first four progress monitoring tests and pp3 scores at the end of grade 2 were only slightly lower, with $.61 \leq r \leq .68$. Later progress monitoring tests related to the pp2 and pp3 scores to a somewhat lesser degree (see Table 4).

---

[2] Our study design resulted in data with a hierarchical structure (students nested in classrooms), and some intra-class correlations (ICC) suggested that error variances may be underestimated if this was not accounted for (the mean ICC for all progress monitoring and paper pencil tests was .08). We therefore performed multi-level modelling (using Mplus 7.11) in addition to single-level modelling for all correlational analyses in both studies. Concerning correlations, the maximum absolute difference between the methods in study 1 and 2 was .04 and .03, respectively. The mean difference of all correlation coefficients was <.01 and .01, respectively, with multi-level mean correlations being marginally higher in study 2. Furthermore, there was no meaningful difference in the mean standard error ($M_{\text{diff}} < .01$; the single maximum absolute difference was .03), and all *p* levels were identical. Because of the relatively small number of classrooms and because single-level results are slightly more conservative, we report results from single-level analyses.
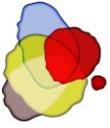
Table 4

Concurrent and predictive validity of progress monitoring scores

Measure	1	2	3	4	5	6	7	8	9	10
1. time 1										
2. time 2	.74									
3. time 3	.70	.80								
4. time 4	.67	.74	.76							
5. time 5	.62	.69	.69	.73						
6. time 6	.64	.67	.74	.77	.73					
7. time 7	.59	.59	.70	.76	.74	.80				
8. time 8	.54	.59	.66	.68	.68	.75	.76			
9. pp1	.41	.50	.47	.44	.45	.47	.43	.40		
10. pp2	.64	.66	.65	.71	.62	.58	.59	.61	.46	
11. pp3[a]	.61	.68	.65	.68	.51	.56	.57	.50	.42	.76

Note. All correlation coefficients were statistically significant at an alpha level of $p < .001$. pp = paper pencil test [a] $n = 148$

3.2.2 Teacher ratings

Teachers' ratings of their students' mathematical ability were correlated with the progress monitoring test scores (see Table 5). Results initially revealed low to moderate correlations between the progress monitoring scores and ratings provided at the beginning of grade 1 (teacher rating 1; $.29 \leq r \leq .42$). Correlations with ratings provided at the end of grade 1 were substantially higher (teacher rating 2; $.54 \leq r \leq .64$) and remained stable for ratings provided at the end of grade 2 (teacher rating 3; $.54 \leq r \leq .66$), indicating high predictive validity.
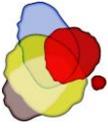
Table 5

Correlations between progress monitoring scores and teacher ratings of students' mathematical ability, provided at grade 1 fall, grade 1 summer, and grade 2 summer

progress monitoring	teacher ratings		
	1	2	3
time 1	.39	.60	.60
time 2	.40	.62	.66
time 3	.42	.64	.66
time 4	.37	.59	.62
time 5	.38	.54	.58
time 6	.34	.60	.59
time 7	.29	.54	.58
time 8	.37	.56	.54

Note. All correlation coefficients were statistically significant at an alpha level of $p < .01$.

3.3   Usability and practicality

Median test time for the first progress monitoring test was 15.48 minutes ($SD = 4.81$). Later test times were considerably lower and declined continuously, from 13.85 minutes for test 2 ($SD = 4.37$) to 8.20 minutes for test 8 ($SD = 3.81$). The difference between the first test and all other tests was partly due to initial starting introductions to the test (approx. 1 minute) and to the students' unfamiliarity with the system.

In the survey about the progress monitoring tests, students rated the tests highly, with mean scores of 4.28 ($SD = 1.05$) on the question, "How did you like the tests?" and 4.34 ($SD = 1.13$) on the item, "Would you like to do the tests again next school year?" (on a smiley faces scale from 1, *very unhappy* to 5, *very happy*). 4% and 7% of the students rated the items negatively (scale points 1 or 2), opposed to 71% and 78% positive ratings (scale points 4 or 5).

The 10 teachers who participated in study 1 gave similar estimations in the questionnaire provided after completion of the progress monitoring tests. On the 4-point Likert scale (*disagree* to *agree*), all teachers agreed that, "most of the students had fun completing the tests" ($M = 3.70$). The same distribution of answers was found for the item, "The students were able to conduct the tests independently". Nine teachers stated that the added benefit of the tool was worth the additional timely effort ($M = 3.10$). Moreover, these teaches stated that they would continue to use the system in the next school year ($M = 3.60$) and recommend the program to fellow colleagues ($M = 3.50$). Teachers declared that they used the progress monitoring results diversely for classroom purposes. Apart from obtaining general performance information at student and class level (100%, 70% agreement, respectively), teachers found the information especially useful when they were previously unsure of a student's performance (70% also used the system for this purpose). Most teachers adjusted their estimate of students' performance for some students (80% agreement) and claimed to have at least sometimes given weaker or stronger students adjusted exercises based on progress monitoring test results (70%, 90% agreement for weaker or stronger students, respectively). Eight teachers stated that supplementary education for weak students was offered at their schools, and information from the progress

monitoring tests was used for designing the supplementary education at six of these schools. A majority of respondents also found the information important for communicating about performances with students, parents and fellow teachers (90% agreement). The main concern of several teachers participating in the study was the two-week time frame per test in that study. They wished for three-week testing intervals to allow more time for analysing and working with the results.

4.   Results study 2

While study 1 evaluated the test's validity as well as its usability and practicality, study 2 focused on the reliability and sensitivity to learning. With respect to the different aims of the two studies, analyses also differed between the studies. Additionally, given the extended test intervals and because some of the test items were adjusted concerning their difficulty for study 2, results differ slightly from study 1.

4.1   Alternate-form reliability

We calculated the delayed alternate-form reliability for each adjacent test (t1 × t2, t2 × t3, … t7 × t8). Coefficients ranged from $r = .71$ to $.83$ ($M = .78$), which is a sign for parallelism across tests. Parallelism is also indicated by the pattern of correlations between non-adjacent tests (see Table 6), which decreased only slightly with increasing amount of time between the probes (e.g., test 1 × test 4).

Table 6

Delayed alternate-form reliability of progress monitoring scores, study 2

progress monitoring	1	2	3	4	5	6	7
1. time 1							
2. time 2	.71						
3. time 3	.65	.74					
4. time 4	.68	.76	.81				
5. time 5	**.67**	.71	.78	.82			
6. time 6	.60	**.60**	.64	.74	.77		
7. time 7	.57	.63	**.67**	.74	.77	.79	
8. time 8	.59	.67	.69	**.69**	.75	.76	.83

Note. Correlations of same test forms are printed in bold. All correlation coefficients were statistically significant at an alpha level of $p < .001$.

4.2   Sensitivity to learning

The test's overall capacity to assess learning gains was determined by calculating growth rates in test scores using linear regression for the eight tests. Weekly growth rates were obtained by dividing the resulting slopes by 3 because of the three-week time frame of each test. Weekly increases in overall scores of 1.0 percent points could be observed (see Table 7; descriptive statistics for study 1 are listed in the appendix),
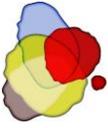
with larger weekly gains for Advanced Precursors and Computation skills than for Basic Precursors. Smaller Basic Precursors gains are mainly due to the Symbolic Quantity Discrimination task which revealed ceiling effects from the first probe (see Figure 2).

Table 7

Descriptive statistics and growth rates for competences, study 2

progress monitoring	overall score		Basic Precursors		Advanced Precursors		Computation	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
time 1	62.1	11.7	79.4	13.0	55.9	19.0	46.1	15.4
time 2	66.5	14.0	79.1	12.6	65.6	21.9	50.6	19.1
time 3	70.8	14.5	83.1	12.4	66.6	23.1	59.1	19.5
time 4	74.3	14.8	84.6	10.9	73.0	23.0	61.9	22.9
time 5	75.6	13.6	86.1	10.5	72.5	20.5	65.0	20.7
time 6	78.1	14.5	87.0	11.2	74.3	20.6	70.6	21.5
time 7	82.8	13.4	90.1	9.6	80.2	21.3	76.0	21.3
time 8	81.2	14.4	88.5	11.1	77.8	22.6	75.2	22.6
Growth rate	1.0		0.5		1.0		1.5	

Note. All scores as percentage correct. Growth rates are weekly growth rates, calculated as slopes of linear regressions of the 8 tests divided by 3 (because of the three-week delay between each test in study 2).
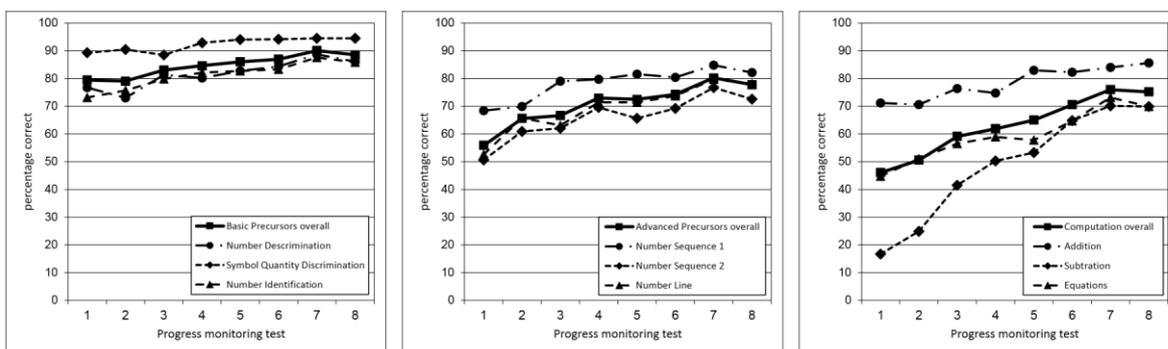

Figure 2. Growth rates for single measures in study 2 ($n = 153$).

Statistical significance of growth rates for overall scores was examined by conducting repeated-measures analyses of variance. Mauchly's Test revealed a violation of sphericity ($p < .001$). Thus, Greenhouse-Geisser corrections were used (Greenhouse & Geisser, 1959). Results indicate an effect of time, $F(5.50, 836.18) = 137.73$, $p < .001$, $\eta^2 = .48$. There was also a significant effect of time for the three single
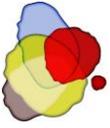
competences Basic Precursors, $F(6.22, 945.13) = 35.14$, $p < .001$, $\eta^2 = .19$; Advanced Precursors, $F(6.04, 917.67) = 51.47$, $p < .001$, $\eta^2 = .25$; and Computation, $F(5.63, 855.82) = 96.95$, $p < .001$, $\eta^2 = .39$. Post hoc tests were performed to analyse for significant increases from test to test. All six increases in total scores from test 1 to test 7 were significant (see Table 8). However, scores decreased from test 7 to test 8. For Basic Precursors and Advanced Precursors, 4 and 3 of the six increases from test 1 to 7, respectively, were significant ($p < .05$) as well as all six increases for Computation scores. Decreases from test 7 to 8 were significant only for Advanced Precursors, $t(152) = 1.69$, $p = .049$.

Table 8

Comparisons of mean differences in progress monitoring scores for study 2

comparisons	mean score difference (*SD*)	*t*	*df*	*P*
time 1 – time 2	-2.26 (5.20)	-5.37	152	< .001***
time 2 – time 3	-2.25 (5.32)	-5.23	152	< .001***
time 3 – time 4	-1.80 (4.73)	-4.72	152	< .001***
time 4 – time 5	-0.67 (4.46)	-1.87	152	.031*
time 5 – time 6	-1.33 (5.01)	-3.28	152	.001**
time 6 – time 7	-2.45 (4.77)	-6.18	152	< .001***
time 7 – time 8	0.86 (4.23)	2.24	152	.014*

5.   Discussion

The current study extends the research on progress monitoring for young students by using an automated assessment tool that allows frequent tests in regular-education settings and provides educators with detailed information about students' skills. The primary goal of the study was to determine the adequacy of the newly-developed progress monitoring tool. First-grade students work independently on the short online tests, so that diagnostic information about students' performance and progress is obtained with minimal instructional time. The tool uses a combination of robust indicator and curriculum sampling approaches to comprehensively assess nine short measures of mathematic performance forming three competences. Static scores and longitudinal psychometric properties were investigated alongside feasibility and usefulness for instructional changes.

First, with regard to reliability, the overall scores of the progress monitoring tests showed good internal consistencies within a narrow range. Consistencies of individual competence scores – particularly Basic Precursors and Computation – were considerably lower. Low coefficients for Basic Precursors may be due to ceiling effects; Computation consistencies were larger for later tests, which may indicate that the three measures within the competence set are distinct skills at first. The distribution of difficulties (see Figure 2) contributes to this interpretation. Correlations between adjacent tests as a measure of delayed alternate-form reliability were strong, which indicates reliable assessment of students' performance despite the young age of the students. Increasing adjacent-test correlations after test 3 (see Table 6) argue that frequent tests are advantageous.

Second, progress monitoring tests 1 to 4 were closely related to the paper pencil results and teacher ratings at the end of first and second grade (pp2 and pp3). Noteworthy is the stability of the predictions over time, which indicates that the progress monitoring tests in the first half of the school year assess skills
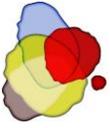
particularly important for long-term mathematics success. Somewhat lower correlations between tests 5 to 8 and the standardised tests pp2 and pp3 may be because – as indicated in Figure 2 – some children showed ceiling effects at the end of the school year. Some ceiling effects are a desired result because test items are designed to represent end-of-year competence goals, which several students typically already reach earlier in the school year. Yet, reduced variance of progress monitoring tests is likely to result in a slight reduction of correlations with standardised measures of mathematical competence.

Progress monitoring results were less closely related to the paper pencil test at the beginning of grade 1, which merely assessed precursor abilities and was only moderately predictive of the results of the later paper pencil achievement tests (see Table 4). Moderate predictive value was also observed for the first performance ratings by the teachers, who had known their students for about two months at that time (correlations between teacher rating 1 and pp2/pp3 were $r = .44$ and .43, respectively). Thus, in addition to the detailed results on precursor abilities from standardised tests (e.g., OTZ), the progress monitoring tests can provide teachers with information about students' abilities vital for long-term learning growth.

Third, the tests proved to be sensitive to learning growth with increasing scores from progress monitoring test 1 to 8 in all competences. However, some scores decreased in the last test, an occurrence which has also been observed in other progress monitoring research when frequent tests were conducted (Förster & Souvignier, 2011; Hampton et al., 2012). For progress monitoring tests 1 to 7, all test-to-test increases were significant for overall scores and Computation. For Basic Precursors and Advanced Precursors – skills that were expected to be mastered before or soon after school entrance – higher overall scores than for Computation were observed, and only some of the increases were significant. Thus, growth patterns of these two single competences should be interpreted with caution and over longer time periods.

Finally, several measures of feasibility and usefulness of the tool showed adequate results. The time that students needed to complete a test was low, and the students were able to work on the tests independently. The remaining implementation effort was justified in the eyes of the teachers, a precondition for frequent and beneficial use. Teachers also stated that they used the results in diverse ways for classroom purposes and individualised instructions, although the exact scope of instructional changes remains unknown.

To conclude, the study at hand addresses a number of issues that were discussed in previous research. By including measures from two approaches, robust indicators and curriculum sampling, the progress monitoring tool provided teachers with performance information about tasks which are directly related to classroom work. At the same time, the combination of different measures proved to be reliable and highly predictive of students' short- and long-term performance. Overall scores increased from test to test for all but the last data point, enabling teachers to judge their students' progress and implement necessary interventions rapidly. Low testing times and concise results views provide an adequate basis for use in general education.

5.1   Limitations

At least five limitations should be taken into account when generalising the findings of this study. First, although the participating classrooms were selected from rural and urban areas in different school districts, all schools were in the same federal state, and results could differ in other regions of Germany.

Second, the differing test intervals and slightly adjusted test items between study 1 and 2 limit the comparability of results between the studies.

Third, no direct measure of parallel-forms reliability was obtained because different test forms were not administered at the same time. All test items were designed using detailed algorithms to ensure similar difficulties, and narrow-ranging reliability coefficients (a) for adjacent tests in study 2 and (b) for predictive validity in study 1 suggest some degree of parallelism. Nonetheless, parallelism of the test concept should be assumed with caution until direct parallel-forms reliability has been determined.

Fourth, slightly larger test score increases in the first few progress monitoring tests (when students are still somewhat unfamiliar with the computer tests) may indicate some degree of retest effects. However,
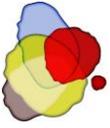
large differences in the slopes of different measures (cf. Figure 2) and teachers' ratings of the usability of the tests for children suggest that this effect is small.

Finally, the added value of the Basic Precursors competence for the majority of students remains questionable. Basic Precursors scores showed ceiling effects early, with low internal consistencies and limited increases over time. The competence was included in the test as a measure for skills which students should already have acquired before school entrance. Teachers should therefore pay special attention to students who do not reach high Basic Precursors scores.

5.2   Implications for research and practice

Several different competences were included in the test concept at hand to provide teachers with detailed information about students' strengths and weaknesses, as recommended by Methe (2012). Overall scores were highly predictive of the students' long-term learning outcome, and teachers stated to utilise the information for individualised instruction and supplementary education. Single competence scores in part showed lower levels of internal consistency and sensitivity to learning growth than desired. Teachers should thus prefer overall test scores when making high-stakes educational decisions. Results of the nine single measures can be used at individual level to detect specific deficiencies that prevent a student from advancing in other competence areas. All in all, general education teachers can use the progress monitoring tool to reliably and quickly assess different aspects of their students' mathematics performance and the development over time. A review by Stecker et al. (2005) showed that the use of progress monitoring tools resulted in higher learning gains specifically if educators were provided with diverse information about student competences, which they then utilised for individualised instruction. Most participating teachers in our study stated that they used the results to adjust their classroom work. However, the extent and success of these adjustments have not been assessed.

We recommend two fields of interest for further research in this domain. First, the specific contribution of single competences for the performance of different groups of students remains to be determined. For low-performing students, certain precursor cut-off scores may provide a more accurate risk estimation of long-term mathematics success than total scores. Second, it remains largely unexplored how teachers systematically use progress monitoring information to enhance student learning. Although the tool at hand includes several measures that are directly related to the curriculum, the review by Stecker et al. (2005) suggests that teachers need additional support with "translating" diagnostic information into improved classroom work.

Keypoints

- Web-based progress monitoring is used for highly automated documentations of learning progress
- Scores of progress monitoring tests are highly predictive of mathematics performance at the end of first and second grade
- First-grade students worked on the tests independently and with high satisfaction
- The short tests with nine different measures in three competences were sensitive to learning growth, showing test-to-test increases
- Teachers stated to use progress monitoring results diversely for individualised instruction
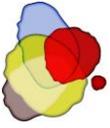
References

Aunola, K., Leskinen, E., Lerkkanen, M.-K., & Nurmi, J.-E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology*, *96*(4), 699–713. doi:10.1037/0022-0663.96.4.699

Baglici, S. P., Codding, R. S., & Tryon, G. (2010). Extending the research on the tests of early numeracy: Longitudinal analyses over two school years. *Assessment for Effective Intervention*, *35*(2), 89–102. doi:10.1177/1534508409346053

Berch, D. B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities*, *38*(4), 333–339. doi:10.1177/00222194050380040901

Chard, D. J., Clarke, B., Baker, S. K., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention*, *30*(2), 3–14. doi:10.1177/073724770503000202

Clarke, B., Baker, S., Smolkowski, K., & Chard, D. J. (2008). An analysis of early numeracy curriculum-based measurement: Examining the role of growth in student outcomes. *Remedial and Special Education*, *29*(1), 46–57. doi:10.1177/0741932507309694

Clarke, B., Nese, J. F. T., Alonzo, J., Smith, J. L. M., Tindal, G., Kame'enui, E. J., & Baker, S. K. (2011). Classification accuracy of easyCBM first-grade mathematics measures: Findings and implications for the field. *Assessment for Effective Intervention*, *36*(4), 243–255. doi:10.1177/1534508411414153

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, *33*(2), 234–248.

Clause, C. S., Mullins, M. E., Nee, M. T., Pulakos, E., & Schmitt, N. (1998). Parallel test form development: A procedure for alternate predictors and an example. *Personnel Psychology*, *51*(1), 193–208. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=487650&lang=de&site=ehost-live

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. doi:10.1037/1082-989X.6.4.330

Connor, C. M., Morrison, F. J., & Petrella, J. N. (2004). Effective reading comprehension instruction: Examining child x instruction interactions. *Journal of Educational Psychology*, *96*(4), 682–698. doi:10.1037/0022-0663.96.4.682

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44*(1-2), 1–42. doi:10.1016/0010-0277(92)90049-N

Dehaene, S. (2011). *The Number Sense. How the mind creates mathematics* (2nd ed.). New York, NY: Oxford University Press.

Dehaene, S., & Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathematical Cognition*, *1*(1), 83–120.

Deno, S. L. (2003). Curriculum-based Measures: Development and perspectives. *Assessment for Effective Intervention*, *28*(3-4), 3–11. doi:10.1177/073724770302800302

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., … Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. doi:10.1037/0012-1649.43.6.1428

Foegen, A., Jiban, C. L., & Deno, S. L. (2007). Progress monitoring measures in mathematics. *The Journal Of Special Education*, *41*, 121–139.

Förster, N., & Souvignier, E. (2011). Curriculum-based measurement: developing a computer-based assessment instrument for monitoring student reading progress on multiple indicators. *Learning Disabilities: A Contemporary Journal*, *9*(2), 65–88.

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities*, *38*(4), 293–304.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112.
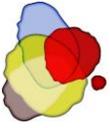
Hampton, D. D., Lembke, E. S., Lee, Y.-S., Pappas, S., Chiong, C., & Ginsburg, H. P. (2012). Technical adequacy of early numeracy curriculum-based progress monitoring measures for kindergarten and first-grade students. *Assessment for Effective Intervention*, *37*(2), 118–126. doi:10.1177/1534508411414151

Jordan, N. C., Kaplan, D., Oláh, L. N., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development*, *77*(1), 153–175. doi:10.2307/3696696

Kim, Y.-S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology*, *102*(3), 652–667. doi:10.1037/a0019643

Koponen, T., Aunola, K., Ahonen, T., & Nurmi, J.-E. (2007). Cognitive predictors of single-digit and procedural calculation skills and their covariation with reading skill. *Journal of experimental child psychology*, *97*(3), 220–41. doi:10.1016/j.jecp.2007.03.001

Krajewski, K. (2008). Prävention der Rechenschwäche. [The early prevention of math problems]. In W. Schneider & M. Hasselhorn (Eds.), *Handbuch der Pädagogischen Psychologie* (pp. 360–370). Göttingen: Hogrefe.

Krajewski, K., Küspert, P., & Schneider, W. (2002). *DEMAT 1+. Deutscher Mathematiktest für erste Klassen. [German mathematics test for first grades]*. Göttingen: Beltz Test.

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

Krajewski, K., & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction*, *19*(6), 513–526. doi:10.1016/j.learninstruc.2008.10.002

Methe, S. A. (2012). Innovations and future directions for early numeracy curriculum-based measurement: Commentary on the special series, part 2. *Assessment for Effective Intervention*, *37*(2), 67–69. doi:10.1177/1534508411431256

Methe, S. A., Begeny, J. C., & Leary, L. L. (2011). Development of Conceptually Focused Early Numeracy Skill Indicators. *Assessment for Effective Intervention*, *36*(4), 230–242. doi:10.1177/1534508411414150

Missall, K. N., Mercer, S. H., Martínez, R. S., & Casebeer, D. (2012). Concurrent and longitudinal patterns and trends in performance on early numeracy curriculum-based measures in kindergarten through third grade. *Assessment for Effective Intervention*, *37*(2), 95–106. doi:10.1177/1534508411430322

Newton, H. J., Baum, C., Clayton, D., Franklin, C., Garrett, J. M., Gregory, A., … Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, *4*(3), 227–241. Retrieved from http://www.stata-journal.com/article.html?article=st0067

Rubin, D. B. (1987). *Statistical analysis with missing data* (4th ed.). New York, NY: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18 + years. *Journal of the American Statistical Association*, *91*(434), 473–489. doi:10.1080/01621459.1996.10476908

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. doi:10.1037/1082-989X.7.2.147

Seethaler, P. M., & Fuchs, L. S. (2011). Using curriculum-based measurement to monitor kindergarteners' mathematics development. *Assessment for Effective Intervention*, *36*(4), 219–229. doi:10.1177/1534508411413566

Siegler, R. S., & Booth, J. L. (2004). Development of Numerical Estimation in Young Children. *Child Development*, *75*(2), 428–444. doi:10.1111/j.1467-8624.2004.00684.x

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: review of research. *Psychology in the Schools*, *42*(8), 795–819. doi:10.1002/pits.20113

Van Luit, H., van de Rijt, B., & Hasemann, K. (2001). *Osnabrücker Test zur Zahlbegriffsentwicklung [Osnabrück test of number concept development]*. Göttingen: Hogrefe.