Department of Psychology and Sport and Exercise Sciences

Institute of Psychology in Education

# Mathematics Progress Monitoring in the Primary Grades:

## Construction and Validation of Progress Monitoring Test Concepts in Grade 1 and 2

Inaugural Dissertation
for Obtaining a Doctoral Degree
at the Department of Psychology and Sport and Exercise Sciences,
University of Münster

by

*Martin Clemens Salaschek*

born in Hamburg

– 2014 –

*Graduation committee*

1. Prof. Dr. Elmar Souvignier
2. Prof. Dr. Regina Jucks
3. Dr. Jörg-Tobias Kuhn

| | |
|---|---|
| Dean: | Prof. Dr. Manfred Holodynski |
| First reviewer: | Prof. Dr. Elmar Souvignier |
| Second reviewer: | Prof. Dr. Regina Jucks |
| Day of oral examination: | 17 February 2014 |
| Day of graduation: | ____________________________ |

# Manuscripts

Salaschek, M., & Souvignier, E. (accepted). Web-based progress monitoring in first grade mathematics. *Frontline Learning Research*.

Salaschek, M., & Souvignier, E. (under review). Web-based mathematics progress monitoring in second grade. *Journal of Psychoeducational Assessment.*

Salaschek, M., Zeuch, N., & Souvignier, E. (under review). Mathematics growth trajectories in first grade: Cumulative vs. compensatory patterns and the role of number sense. *Learning and Individual Differences.*

# Summary

Progress monitoring tests provide teachers with diagnostic information about student performance. This information can then be used to enhance classroom work. With this dissertation, I aimed to create and validate computer-based progress monitoring test concepts which provide teachers with information about their students' mathematics performance at a single time point and over the course of time. In the course of the studies, psychometric properties of the tests as well as the feasibility of implementation were evaluated, and learning trajectories of students with diverse prior knowledge were explored.

I drew on models describing the development of mathematics competences early in life to develop computer-based test concepts which comprehensively assess mathematics competences in grade 1 and 2. Key curricular computation tasks complemented the tests for increased criterion validity and interpretability. For each grade, parallel test forms were created which can be administered in short intervals throughout a school year.

Two manuscripts describe the progress monitoring implementation in general-education settings and explore the tests' psychometric quality. In both manuscripts, correlations of adjacent tests indicated reliability of the assessments. School achievement tests as well as teacher ratings of their students' mathematics competence were used to explore criterion validity. Strong correlations were found particularly for longer-term performance predictions. Moreover, significant linear increases in absolute scores suggest that the tests reliably depict learning growth. Teachers declared that implementation of the tests was feasible and that the obtained diagnostic information was useful for classroom work.

First-graders' diverse development patterns were explored in a study described in the third manuscript. Results emphasize the importance of precursor competences for acquiring grade-level curricular skills. While students followed mainly fan-spread performance patterns—i.e., initially low-performing students showed less learning growth over the school year than initially high-performing students—there were some students in all competences with initially low scores but steep learning growth. Students with such a performance pattern in precursor competences did not have elevated risks of performing poorly in higher-level skills by the end of the school year.

In conclusion, the newly-developed tests provide teachers with reliable diagnostic information about their students' competences. Further research is needed on how teachers can use this information for increased student learning.

# Contents

# PART I

Introduction and Discussion

# 1

# Introduction

## 1.1 The demand for mathematics progress monitoring in primary school

Teachers face great individual competence differences in the classroom: Even at the beginning of first grade, children's previous knowledge in mathematics and reading varies extensively (Aunola, Leskinen, & Nurmi, 2006; Jordan, Kaplan, Oláh, & Locuniak, 2006; Morgan, Farkas, & Wu, 2011; Parrila, Aunola, Leskinen, Nurmi, & Kirby, 2005). Longitudinal studies provide evidence that these performance differences are stable over time or even grow larger over several years. In other words, students with weak initial performance generally seem to fall increasingly behind (Bast & Reitsma, 1997; Bodovski & Farkas, 2007; Jordan, Kaplan, Locuniak, & Ramineni, 2007; Leppänen, Niemi, Aunola, & Nurmi, 2004; Lerkkanen, Rasku-Puttonen, Aunola, & Nurmi, 2004; Parrila et al., 2005).

It also is important to note, though, that effective interventions for low-performing students do exist and that they seem to be most advantageous if they are conducted early, so that deficits do not cumulate (Bryant, Bryant, Gersten, Scammacca, & Chavez, 2008a, 2008b; Bus & van IJzendoorn, 1999; Dyson, Jordan, & Glutting, 2013; Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Fuchs, Fuchs, & Compton, 2012; Gersten, Jordan, & Flojo, 2005; Sood & Jitendra, 2013). When whole classrooms are of interest instead of just low-performing students, several studies suggest that children at low through high performance levels generally profit from instruction that is adjusted to student characteristics (Connor, Morrison, & Petrella, 2004; Freebody & Tirre, 1985; McDonald Connor et al., 2009; Stecker & Fuchs, 2000).

Consequently, teachers should strive to identify strengths and weaknesses of their students and modify their classroom work accordingly—a demand which is also increasingly expressed in politics (e.g., § 1 *Schulgesetz für das Land Nordrhein-Westfalen*, 2013). While teachers have a wide range of possibilities to diagnose student performance characteristics (see Förster, 2013, for a discussion), the accuracy of teachers' performance judgments differs largely (Hoge & Coladarci, 1989; Lorenz & Artelt, 2009), and estimations are less accurate for low-performing students (Feinberg & Shapiro, 2009).

Screening tools can aid teachers in obtaining more exact estimations of achievement levels and students' risks of failing the curricular goals, but Gersten et al. (2012) discussed several unanswered questions concerning specificity and sensitivity of the tools. One of the concerns raised by the authors is that some students may show sufficient performance at first but fail to learn at an acceptable rate later on, and other students with poor early performance may show compensatory patterns with narrowing achievement gaps over time. While the former assumption is supported by some findings in reading (Scarborough, 2009) as well as mathematics for older students (Ehmke, Blum, Neubrand, Jordan, & Ulfig, 2003), convincing results of narrowing patterns for the majority of low-performing students have only been found in reading to date (see Morgan et al., 2011, for an overview). However, Morgan, Farkas, and Wu (2009) found that only about half of the students scoring in the lowest decile in a standardized mathematics achievement test in the fall of kindergarten still scored in the lowest decile in the spring of kindergarten[1]. It thus seems advisable to systematically monitor students' performance development over time to improve performance classifications (Gersten et al., 2012).

In sum, previous research suggests that (1) students enter school with large differences in curricular competences; (2) initially low-performing students tend to fall increasingly behind; (3) interventions for low-performing students are particularly promising if implemented early; (4) also students on other performance levels profit from individualized instruction; (5) systematic monitoring of students' progress is advisable, given students' diverse performance development and the resulting issues with risk estimations of traditional screening tools (administered at only one time point).

Several of these issues have been studied less intensely in mathematics than in reading. Therefore, the objective of the present dissertation was to construct mathematics test concepts which inform general-education teachers about their students' performance longitudinally, beginning with the start of formal schooling.

[1] In the US, where the study was conducted, "*kindergarten*" usually refers to a preschool institution, lasting one year before formal school entrance. In the dissertation, this nomenclature is adopted.

## 1.2 Approaches to longitudinal assessments of early mathematics competences

### 1.2.1 Development of mathematics competences

Developmental models of early skills provide insight into what capabilities are needed for the development of *number sense* (see Berch, 2005, for a discussion of the term) and, subsequently, proficiency in primary-grade mathematics. These developmental models thus provide important implications which skills to consider in a longitudinal assessment. In reading, demarcated precursors for reading comprehension have been identified, e.g., phonological awareness, print knowledge, and reading fluency (Kim, Petscher, Schatschneider, & Foorman, 2010; Storch & Whitehurst, 2002). In mathematics, precursor competences associated with later math achievement are more diverse, and developmental models describe several number and magnitude skills which start to develop early in life.

Stanislas Dehaene and colleagues have contributed a large body of research in this field, including the repeatedly extended and revised *triple-code model of number processing* (Dehaene, 1992, 2001, 2011; Dehaene & Cohen, 1995). The triple-code model does not provide detailed suggestions about the time or order of mathematical skill development. Nonetheless, it is well-suited for a deduction of what skills constitute number sense. The model depicts three distinct systems involved in number processing: (1) a quantity system or *analogue magnitude representation* (with nonverbal semantic representations of size and distance relations), (2) a verbal system or *verbal word frame* (verbal representations of numerals), and (3) a visual system or *visual Arabic number form* (for written numerals, usually Arabic). The model proposes that the three systems develop independently. The meaning of quantities and numbers is located solely in the quantity system, on an "oriented number line" (Dehaene & Cohen, 1995, p. 86). Verbal representations and the knowledge of Arabic numbers do not convey meaningful quantity information themselves. The three systems, however, usually interact with each other, and information is transcoded along the way.

Krajewski built on Dehaene's findings in her *model of early mathematical development* (Krajewski & Schneider, 2009; Krajewski, 2008). This model takes a more sequential frame of reference than the triple-code model and describes three levels of early mathematical competences which serve as milestones for the development of number sense and mathematical thinking. These theoretical considerations were confirmed by the authors in a four-year longitudinal study, where precursor skills on the first two levels were particularly predictive of mathematical achievement in primary school (Krajewski & Schneider, 2009).

The first level in the model, *Basic numerical skills*, consists of (a) basal quantity discrimination skills and (b) the acquisition of number words and subsequently the exact number-word sequence for small numbers. Skills on this level progressively develop with the acquisition of language (Krajewski & Schneider, 2009, p. 514). In the second level, *Linking number words with quantity*, number words are filled with meaning, as described in the triple-code model. When concrete quantities are gradually linked to their precise verbal representation (e.g., "*three* apples"), comparing the magnitude of number words becomes attainable. Skills on this level usually advance between the age of three and five. In the third level, *Linking quantity relations with number words*, children learn to compose and decompose numbers ("five" can be divided into "three and two") and to denote the difference between two magnitudes with a number ("five is *two more* than three"). Thus, a basis for addition and subtraction skills is being laid out.

Krajewski's model illustrates how quickly numerical competences develop in early childhood. However, this development is not simultaneous for all numbers, and the competence levels in Krajewski's model are each reached for small quantities at first. Therefore, Krajewski and Schneider (2009) note that children may operate on different skill levels with small and large numbers. Children may have reached level three of the model for quantities up to five, but may still be on level one for numbers larger than 20.

Krajewski's model does not specify the development of curricular competences in school. Yet, the advancement of skills defined in common curricula still develop at a high pace: For the computational domain alone, in first grade, magnitudes and number words need to be connected to their Arabic representations, arithmetic symbols (+, -) are studied and formal computation in the number range of 1 to 20 is established, including crossing the tens boundary. Basic computation skills are then extended in second grade, when the number range is stretched to 100, multiplication is exercised, and first concepts of division are introduced (e.g., NCTM, 2012; Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, 2008).

In sum, the development of math competences is diverse. Numerous precursor skills should be acquired in the preschool age, at least for small quantities. With the beginning of formal schooling, these precursor competences need to be transferred to larger number ranges to form the basis for first-grade curricular competences. For monitoring students' competence development early in school, this means that several precursor and curricular skills with different levels of complexity should be assessed. The fast rate of increasing curricular demands proposes that students' progress should be observed regularly. This way, students' weaknesses or low learning growth can be identified and classroom work can be adjusted accordingly or further interventions put in place.

### 1.2.2 Progress monitoring in mathematics: Curriculum-based measurement

The goal of continuously checking student achievements has a decade-long history in the US, where, with curriculum-based measurement (CBM), a form of progress monitoring assessment was introduced in the 1970s (Deno, 1985). In CBM, short tests are administered frequently (e.g., weekly). Each test is to be parallel, i.e., the same tasks with the same level of difficulty are included in all tests. A change in test scores therefore represents a change in skill, if reliability of the assessment is high. By that, educators can verify whether students are increasing their skill level at a satisfactory rate (Fuchs, 2004).

CBM was originally introduced in special education as a means to assess learning growth of ongoing curricular activity, and the assessments initially comprised changing curricular content as the school year progressed. Most of today's CBM tests instead determine the progress of the same single skill over the complete school year. Fuchs, in her seminal work, calls this a "robust indicator" approach, where the task in use "correlates robustly (...) with the various component skills constituting the academic domain" (Fuchs, 2004, p. 189). CBM tests following this approach have been shown to determine students' performance very reliably, and growth throughout the school year has regularly been reported (see Foegen, Jiban, & Deno, 2007, for an overview in mathematics). Assessing only one skill has the additional advantage that assessment times can be very short.

### 1.2.3 Curriculum-based measurement in early mathematics education

In mathematics CBM for first grade and higher, most tests require students to solve as many grade-level computation problems as possible within a given time limit (usually one or two minutes). Tests can be group-administered, and the teacher scores each test sheet;

the test result is calculated as the total number of correct digits in the solutions. Tests using this robust indicator approach often yield reliability scores of .80 and higher (Foegen et al., 2007). However, Christ, Scullin, Tolbize und Jiban (2008, p. 204) argue that there is lacking evidence of construct validity for this approach and that mere assessments of computation skills "should not be interpreted to represent mathematics achievement generally". Moreover, single-skill results per se do not allow conclusions about specific strengths or weaknesses. If progress monitoring is to be used for individual adjustments of classroom instructions for students of all performance levels, more specific information about students' skills is required. While additional assessments over and above progress monitoring results deem feasible in special education (where CBM is still mainly used today), time constraints in general education usually do not allow elaborate diagnostics—even less so for all students in a classroom on a regular basis, which would be needed for continuously individualized instruction for all students.

Recently, CBM tests for kindergarten and first grade have been developed in mathematics which use a more diverse approach. Test concepts for this age mostly comprise four tasks to assess number sense: *oral counting*, *number identification*, *quantity discrimination*, and *missing number*. These tasks, called *tasks of early numeracy* (TEN), are administered individually (e.g., to assess the number of mistakes in a counting sequence), and the test result is typically scored separately for each measure. TEN tasks have also demonstrated adequate levels of reliability and predictive value for later mathematics achievement in a number of studies during kindergarten and first grade (e.g., Baglici, Codding, & Tryon, 2010; Chard et al., 2005; Clarke & Shinn, 2004; Missall, Mercer, Martínez, & Casebeer, 2012), although psychometric results for the single tasks vary from study to study, questioning the common practice to interpret scores from each of the four tasks separately (Missall et al., 2012). Additionally, given that TEN tasks merely target precursor abilities which do not relate closely to school curricula, their capacity to monitor students' progress of skills relevant for the classroom is limited (Methe, 2012). Relevance for classroom purposes is important because Stecker, Fuchs, and Fuchs (2005) emphasize that improved learning growth from a CBM context can only be expected if teachers use the CBM results for analyses of relevant student skills and adjust their instructions accordingly.

Another unresolved issue in CBM concerns the reliability of growth. With two exceptions (Hampton et al., 2012; Seethaler & Fuchs, 2011), recent studies in mathematics CBM used only two or three assessments per school year to estimate the tests' sensitivity to student learning. Median or mean increases between time points or linear regression slopes are divided by the number of weeks between the assessment to obtain estimates of weekly growth rates (e.g., Methe et al., 2011; Hampton et al., 2012; Seethaler & Fuchs, 2011), and these growth rates are commonly used to evaluate the learning progress of individual students. Ardoin, Christ, Morena, Cormier, and Klingbeil (2013) discussed the reliability of growth estimates in reading CBM and the appropriateness of using the coefficients for progress evaluations in a sophisticated review of the literature. The authors point out that the variance in slope estimates in reading CBM is usually extensive—as is the case in mathematics CBM—and question the use of slope estimates for evaluation of individual student's progress and subsequent high-stakes decisions.

### 1.2.4 Mathematics progress monitoring applications in Germany

Although CBM has a long history of research and application in the US, mathematics progress monitoring with parallel tests is widely unknown in Germany or other German-speaking countries; to my knowledge, only one other peer-reviewed study has been published to date (Strathmann & Klauer, 2010). In the study, the authors describe the use of paper-pencil computation problems for primary-grade students (mostly grade 2 to 4) which was similar to traditional CBM scenarios: The paper-pencil tests were group-administered, and teachers summated all correct answers to a single raw score. In an effort to keep students from copying from each other while ascertaining parallelism of the test forms, the authors defined several attributes that contribute to the difficulty of grade-level computation problems. With these attributes, stratified random sampling was used to create student-individual test forms with 24 computation problems. Tests were administered every two weeks and were untimed.

The authors report narrow-ranging adjacent-test correlations for the two-week interval ($.72 \leq r \leq .81$, $M = .77$). Given the fairly large number of problems in the test and the short test intervals, these values can be seen as fair to satisfactory. Test scores mostly increased significantly over time, but growth results need to be interpreted with caution because of the low number of students per grade level. The study did not include any other aspect of mathematics competence than basic arithmetic operations, neither were measures of criterion validity reported. Practicality for general-education use seems limited because teachers were required to score each student's test per hand, which is time-consuming.

### 1.2.5 Implications for progress monitoring in general education

For a beneficial use of progress monitoring in general education, some characteristics need to be observed which call for new assessment concepts compared to typical CBM tests.

First, implementation of the assessment concept needs to be highly economic in order to accommodate the limited instructional time available for assessment purposes in general education. Face to face assessments as well as manual scoring of results, both common practice in most CBM application scenarios, seem infeasible.

Second, the test concept must display high criterion and curricular validity. Namely, test scores need to correlate with current curricular requirements and be predictive of medium-term and long-term achievement. Test concepts should include precursor skills (to aid risk estimations) as well as tasks which relate closely to the curriculum to facilitate instructional adjustments. Considering that student abilities vary both cross-sectionally and over time, tasks need to comprise a wide range of difficulty levels.

Third, the test needs to be sensitive for changes in students' performance so that the growth in student scores over time provides teachers with reliable information about students' learning progress (or the lack thereof).

## 1.3 Aims of the dissertation

The main aim of this dissertation was to create progress monitoring test concepts which can be used to follow students' development of math skills at the beginning of formal education in first and second grade. In detail, implications from CBM research stretched out the following requirements: (1) The test concepts should display high concurrent and predictive criterion validity; (2) Tests should be sensitive to student learning, i.e., changes in students' competences should be reflected in test scores. For changes in test scores to be interpreted as changes in student competence, the tests need a high level of reliability and parallelism; (3) Performance level and growth in precursor and curricular skills should be explored as to their role in the development of learning trajectory groups. By this, the research basis for data-driven decision making from progress monitoring results was to be strengthened; (4) Implementation of the assessment procedure and obtaining results should be as effortless as possible and feasible for general education classrooms (i.e., fitting the demands and being doable under the constraints of school settings).

# 2
# Test concepts and summary of findings

The aims of the dissertation were pursued within three empirical research manuscripts. The first manuscript, "Web-based progress monitoring in first grade mathematics" (Salaschek & Souvignier, accepted), introduces the newly-created multiple-skill progress monitoring tool and its application at the very beginning of formal schooling. Validity, reliability and sensitivity to learning growth are explored, and feasibility of the implementation for teachers and students ascertained. The second manuscript, "Mathematics growth trajectories in first grade: Cumulative vs. compensatory patterns and the role of number sense" (Salaschek, Zeuch, & Souvignier, under review), explores growth trajectories among students with diverse skill levels at the beginning of first grade. The third manuscript, "Web-based mathematics progress monitoring in second grade" (Salaschek & Souvignier, under review), presents findings from the extended progress monitoring application in second grade, for which a new test concept was created. Psychometric properties as well as feasibility findings are reported.

Three longitudinal research projects form the basis of the studies. In all projects, students completed the computer-based progress monitoring tests during regular classroom hours and without additional support by teachers. A total of eight tests was administered every two or three weeks, so that the online assessments lasted from fall to spring of a school year. Given the limited amount of math problems with similar difficulty levels in the number range of the first two grades, four parallel tests A-D were created for both grades, and the four tests were completed twice throughout the school year (sequence A-D, A-D). In two of the projects, the progress monitoring tests were preceded and followed by standardized (paper-pencil) school achievement tests; teachers rated their students'

*Table 1.* Overview of manuscript 1 and manuscript 3 main results

| Manuscript | | N | test interval | *r* adjacent tests | pm × paper-pencil | | pm × teacher ratings | |
|---|---|---|---|---|---|---|---|---|
| Salaschek & Souvignier (accepted) | Study 1 | 220 | 2 weeks | $.73 \le r \le .80$ ($M = .76$) | OTZ: | $.40 \le r \le .50$ ($M = .45$) | Grade 1 fall: | $.29 \le r \le .42$ ($M = .37$) |
| | | | | | DEMAT 1+: | $.59 \le r \le .71$ ($M = .63$) | Grade 1 spring: | $.54 \le r \le .64$ ($M = .59$) |
| | | | | | DEMAT 2+: | $.50 \le r \le .68$ ($M = .60$)[a] | Grade 2 spring: | $.54 \le r \le .66$ ($M = .60$)[a] |
| | Study 2 | 153 | 3 weeks | $.71 \le r \le .83$ ($M = .78$) | — | | — | |
| Salaschek & Souvignier (under review) | | 414 | 3 weeks | $.81 \le r \le .87$ ($M = .84$) | DEMAT 1+: | $.59 \le r \le .63$ ($M = .62$) | Grade 2 fall: | $.57 \le r \le .61$ ($M = .59$) |
| | | | | | DEMAT 2+: | $.72 \le r \le .77$ ($M = .75$) | Grade 2 spring: | $.64 \le r \le .70$ ($M = .68$) |

*Note.* pm = progress monitoring tests 1-8. All correlations were statistically significant at an alpha level of $p < .001$.

[a] n = 148

overall math competence before each of these tests. Additionally, students and teachers were surveyed about the feasibility as well as further use scenarios of the progress monitoring tool and the results.

For CBM, Fuchs (2004) categorized research into three stages which can also be applied to the studies in this dissertation. Following Fuchs (l.c., p. 189), stage 1 is concerned with "technical features of the static score", namely, psychometric properties of the test score at one time point (cf. the first aim of the dissertation). Research on stage 2 investigates the association between the development of test scores and students' competence development—in other words, whether the test accurately details students' learning progress (cf. the second aim of the dissertation). Studies from the research stage 3 examine whether progress monitoring results are used for instructional adjustments which improve student achievement (compared to a control group). The majority of studies in mathematics CBM deal with stage 1 research, even if multiple assessments were conducted for a study; stage 2 research is reported much less frequently, and only very few studies deal with stage 3 research to date (see Foegen et al., 2007; Methe, 2012; Stecker et al., 2005, for an overview and discussions).

The present dissertation pertains to the first two stages of Fuchs' categorization. In manuscript 1 and 3, the static scores are evaluated with respect to validity and reliability (stage 1), and some general growth characteristics are explored (stage 2; Table 1 provides an overview of the designs and main results from the two manuscripts). In manuscript 2, stage 2 attributes are examined in detail by describing students' diverse learning trajectories.

## 2.1 Test concept creation

The quickly-changing curricular demands and the rapid developments in first and second grade required major differences in the test contents of the two grades. The first-grade test concept comprised three different competences: Given the central role of number sense for further mathematical achievement, several precursor skills were assessed at two levels in this test (*Basic Precursors* and *Advanced Precursors*). In addition, *Computation* tasks assessed relevant curricular skills (addition and subtraction in the range of 1 to 20). This promised the possibility to differentiate diverse performance levels, particularly at the start of the school year, where a detailed assessment of number sense skills provides teachers with important diagnostic information of students' prior knowledge.

In second grade, two competences were included in the test concept: Higher-level *precursor* skills—in an extended number range of 1 to 100—were still included to identify students who are not proficient in these skills after the first school year. Diverse *Computation* problems were included to assess second-grade curricular skills.

The competences in both test concepts were each assessed by multiple *measures* (i.e., types of tasks; see Table 2 for an overview). For computation competences, German state curricula were systematically reviewed and tasks were included that assessed central skills of the school year. This approach had three anticipated advantages. First, it ensured a multi-faceted assessment of the competences at hand, which was expected to lead to high criterion validity. Second, measures with a wider range of difficulty could be included than would have been possible with just one measure. Finally, measures could be included which were not expected to show high criterion validity by themselves, but which were expected to add to the breadth and psychometric quality of a competence while at the same time being closely related to the curriculum.

*Table 2.* Competences and measures included in grade 1 and grade 2 test concepts

| | Grade 1 | | Grade 2 | |
|---|---|---|---|---|
| Competence | Measures | Range | Measures | Range |
| *Precursors* | *Basic Precursors* | | *Precursors (combined)* | |
| | Number Discrimination | 1-100 | Number Recognition | 1-500 |
| | Symbol Quantity Discrimination | 1-10 | Size Comparison | 1-100 |
| | | | Number Line | 1-100 |
| | Number Identification | 1-100 | Axis of Symmetry | — |
| | *Advanced Precursors* | | | |
| | Number Sequence 1 | 1-20 | | |
| | Number Sequence 2 | 1-20 | | |
| | Number Line | 1-20 | | |
| *Computation* | Addition | 1-20 | Addition | 1-100 |
| | Subtraction | 1-20 | Subtraction | 1-100 |
| | Equation | 1-10 | Multiplication | 1-100 |
| | | | Double | 1-100 |
| | | | Divide in half | 1-100 |
| | | | Add up to 100 | 100 |

All direct progress monitoring activity was computer-based: Students used personal login data to access the current test on the project's website. They received instructions to all tasks via headphones, so that task comprehension was not dependent on reading skills. Test items were presented in a multiple choice format with large pictures, and the students clicked on the answer they thought was correct. After students completed a test, their current score and previous total scores were displayed.

Teachers could access students' results separately for precursor and computation competences directly after the test. Results were displayed in graphs and tables at the student level and class level, and additional data points were added as students completed more tests. Mean scores of all participating classes with a surrounding area of one standard deviation could be added to the results view.

## 2.2 Manuscript 1: Mathematics progress monitoring in first grade

Manuscript 1 was based on data from two studies, assessing a total of 373 first-grade students (Table 1). The first study focused on concurrent and predictive criterion validity of the static scores (stage 1), the second study explored reliability and parallelism (stage 1) as well as the tests' capacity to model learning growth (stage 2).

In the first study, students completed the progress monitoring tests in intervals of two weeks from fall to spring. Correlations of the eight tests with the three school achievement tests (fall of grade 1, spring of grade 1, spring of grade 2) confirmed the criterion validity of the tests, with particularly strong predictions of school achievement test results. Grade 2 spring achievement (18 months after the first progress monitoring test) was predicted almost as well as grade 1 spring achievement, although the grade 2 paper-pencil test (DEMAT 2+) assessed grade-level curricular competences that were not included in the progress monitoring tests. Along with equally strong correlations with grade 2 teacher ratings, it can be concluded that the progress monitoring tests assessed comprehensive math competences which show construct validity over and above first grade. As a limitation, the predictive value of the progress monitoring tests was slightly decreased for tests 5-8, which may have been caused by ceiling effects in some of the measures.

The progress monitoring tests were moderately associated with results from the first school achievement test at the beginning of the school year, but this paper-pencil test (OTZ) solely assessed precursor skills. Given that the OTZ scores had lower predictive value for the second and third school achievement tests than progress monitoring precursor scores, the choice of precursor measures in the progress monitoring tests seemed to be well-founded.

In the second study, students completed the progress monitoring tests in intervals of three weeks (from fall to late spring). Single test items were adjusted concerning parallelism for the study. Despite the comparatively long test interval, adjacent-test correlations were strong and thereby indicated reliability.

The tests' capacity to model learning growth was explored via repeated measures analyses of variance, which confirmed linear score increases over time for all competences. Moreover, test-to-test increases of overall scores were observed for test 1 through test 7 (but not for all test-to-test comparisons of single competence scores). The combined result of linear growth and strong adjacent-test correlations also argues for parallelism of the tests.

The third major finding reported in the first manuscript was that teachers and students rated several aspects of the progress monitoring tool highly: Students were able to conduct the tests independently, and assessment times were short. Teachers stated to use the results diversely, e.g., to follow the development of students of whose performance they were previously unsure. This use scenario was reflected in increased judgment accuracy of teachers' performance ratings towards the end of the school year. Finally, teachers confirmed that the use of the progress monitoring tool was worth the time needed for implementation.

## 2.3 Manuscript 2: Growth trajectories in first grade

Manuscript 2, based on data from 153 first-grade students, focused on the diverse developmental patterns of students in first-grade mathematics (stage 2). Latent growth curve modeling (LGCM) confirmed that there was significant variance in students' performance level and growth. Therefore, latent class growth analysis (LCGA) was used to model growth trajectories of students for the three competences and overall scores.

Confirming previous findings in this line of research (cf. section 1.1), mainly fan-spread patterns in scores were found (differences in scores between the students at the beginning of the school year increased throughout the course of the study). However, more complex patterns were found when growth trajectories of the single competences were explored: Students with similarly low initial scores divided into trajectory groups with very different growth over time. This resulted in catch-up patterns that were particularly evident for Advanced Precursors and Computation.

To investigate for preconditions of belonging to favorable or unfavorable outcome groups, the stability of classifications in the trajectory groups across the competences was analyzed. Results from these analyses provided evidence for the important role of precursor skills and for the successful operationalization of precursors in the test concept. In essence, students who did not reach high Basic Precursors scores quickly in the course of the study had a strongly increased risk of also scoring relatively low in Advanced Precursors throughout the study, and students who did not reach high Advanced Precursors scores by the end of the study had a strongly increased risk of belonging to low-performing Computation trajectory groups. Nonetheless, students who started with low precursor scores but then displayed strong learning growth were neither less nor more likely to belong to high- or low-performing Computation trajectory groups.

## 2.4 Manuscript 3: Mathematics progress monitoring in second grade

Manuscript three was based on a study with 414 students who were followed in intervals of three weeks (Table 1). The manuscript targeted the same research questions of validity, reliability, parallelism (all stage 1), sensitivity (stage 2), and feasibility as manuscript 1, but for the second-grade test concept.

To evaluate criterion validity, the progress monitoring tests were again directly preceded and followed by paper-pencil school achievement tests. Correlations of the progress monitoring overall scores with these tests were similar to the correlations found in manuscript 1, albeit slightly higher. As a first sign for reliability of the tests, correlations of test 1-8 with each criterion were also very narrow-ranging. Adjacent-test reliability confirmed this assumption, with correlations that were also slightly higher than in grade 1 and also very narrow-ranging.

Sensitivity to student learning was explored by LGCM in this study, and significant linear growth was again observed. Another result from LGCM was that students differed significantly in their scores at the beginning of the study and that these differences remained mostly stable for number sense and overall scores. Thus, the general second-grade competence development seems to be more constant than in first grade. In contrast to this general finding, LGCM revealed significant differences in students' Computation learning growth. This result suggests the presence of distinct growth trajectories for curricular skills.

Finally, students' and teachers' reports about the feasibility and use of the progress monitoring tool as well as the pattern of teachers' judgment accuracy were very similar to the results in manuscript 1.

# 3
# Discussion

Main objective of the dissertation was to expand the research base in mathematics progress monitoring by developing progress monitoring test concepts for first and second grade general-education classrooms. Psychometric properties of the tests were evaluated and developmental trajectories of first-graders analyzed. Frame of reference for the development of the test concepts and their implementation into classrooms were findings (a) from CBM progress monitoring research and (b) from research on the development of early mathematical competences, or number sense.

## 3.1 Psychometric validation of the test concepts

Two of the dissertational manuscripts, "Web-based progress monitoring in first grade mathematics" and "Web-based mathematics progress monitoring in second grade" explored psychometric properties of the test concepts. Results from these studies confirm that the progress monitoring tests are valid and reliable measures of specific grade-level curricular skills and more comprehensive competences. Although test intervals were considerably longer than they usually are in traditional CBM and despite the inclusion of multiple skills with only a few items each, reliability and validity matched or exceeded the coefficients typically reported in CBM research (see Foegen et al., 2007, for an overview as well as Hampton et al., 2012; Lee, Lembke, Moore, Ginsburg, & Pappas, 2012; Polignano & Hojnoski, 2012; Seethaler & Fuchs, 2011, fore more recent results).

Furthermore, test scores increased significantly over time, which is likely due to increased student abilities (but see limitations below). Finally, implementation of the test concept was well-received by the teachers, and the students were able to work on the tests independently. Yet, the results from both studies leave some questions pending as to specific psychometric properties and the use of the results for classroom purposes.

### 3.1.1 Parallelism

In both studies, no direct measure of parallelism has been obtained because no two test forms were administered at the same time. Several results suggest that the tests are indeed parallel, however: High adjacent-test correlations confirm rank-order stability, notwithstanding the possibility that the overall difficulty of the test forms differed (affecting all students), which would limit interpretability of the absolute scores and increases. Similar levels of difficulty are suggested by linear growth patterns of mean scores across the tests, though, because the same four tests were conducted twice in the sequence A-D, A-D. If the tests had varying difficulty levels, peaks or dips in the development of test scores should have been observed. Nonetheless, the exact degree of parallelism remains unknown until several test forms are conducted at the same time.

### 3.1.2 Sensitivity to student learning

In both grades, students displayed significant growth in all competences and—consequently—in overall scores. In CBM, teachers usually draw on information about average increases in scores to evaluate whether a student's individual growth is adequate (Ardoin et al., 2013). Several preconditions have to be met for this practical application, though. First, as previously discussed, reliability of growth across the assessments needs to be high. Second, the reliability of a student's individual slope also needs to be high enough so that his or her estimated slope value does not differ meaningfully from the true slope. Third, all current recommendations of evaluating student growth assume linear growth in individual and average scores. The assumption of linear individual growth is particularly questionable, as the characteristics of individual growth differs substantially from student to student (e.g., Strathmann & Klauer, 2010).

Christ, Zopluoglu, Monaghen, and Van Norman (2013) conclude from their extensive research with multiple simulation studies that the current practice of using CBM data from a few assessments for instructional changes has little support from empirical evidence. (Note that Christ et al., 2013, used data from reading CBM, but most conclusions probably directly apply to mathematics CBM, too.)

In the present dissertation, the assumption of group-level linear slope and the reliability of slopes were analyzed. While mainly linear slopes were indeed found on the group level (although including quadratic estimates slightly enhanced model fit in grade 2 LGCM analyses), the standard errors of the slopes were substantial, and it was not tested how many assessments would have been necessary for an accurate estimation of study-wide growth.

Throughout the research projects of the dissertation, teachers were not provided with specific advice how to interpret their students' growth patterns, and extensive research is clearly needed in this domain that builds on the findings by Ardoin et al. (2013) and Christ et al. (2013).

### 3.1.3 The role of precursors

Ceiling effects of precursor competences were observed in both grades, particularly for basic precursors in grade 1. Precursor competences were included in the test concepts to assess basic skills preceding the curricular competences, and results from manuscript 2 indicate that they indeed served as gateways for more advanced competences. Thus, precursor scores in the progress monitoring tests at hand (and their development) are particularly valuable for an estimation of students' risk of continued difficulties in math and provide clues about deficiencies that limit learning growth throughout the school year. A similar pattern of high predictive value but ceiling effects was also found by Jordan et al. (2007).

However, some of the tasks proved to be very easy for most of the students, and limited variance due to ceiling effects was observed in these measures from the first test. Omitting such tasks may further improve the psychometric properties of the test. Long-term considerations might include IRT scaling of the test items and, subsequently, providing students only with problems appropriate for their skill level.

### 3.1.4 Dimensionality of the test concepts

Finally, the test concepts used results from several types of tasks to form one competence score. This approach was chosen to achieve high reliability with short assessment times while at the same time including numerous different tasks to obtain a comprehensive picture of a competence. These combined scores displayed good psychometric properties—which were again enhanced by combining the competence scores to overall scores—, and following their development proved useful for risk estimations (cf. results from manuscript 2). Yet, the dimensionality of the competences was not analyzed in detail in the present manuscripts. Results from latent confirmatory factor analyses in grade 2 suggested that the fit of a two-factor model as proposed in the test concept is adequate to very good for the eight tests. Furthermore, the two-factor model fitted the data significantly better than a one-factor model. However, the two factors are also highly correlated, which puts at question whether the competences indeed measure *distinct* skills or rather different *levels* of *closely-related* skills.

In addition, educators may find it harder to use competence scores for skills analyses than scores from single tasks, which are usually reported in early mathematics CBM concepts assessing multiple skills (e.g., Baglici et al., 2010; Methe, Begeny, & Leary, 2011). Subsequent research, exploring ways for teachers to use the present test concepts for enhanced student learning, should examine this issue in detail. The contribution of single skills for a competence at hand should be evaluated, and educators will profit from recommendations as to the importance of different skills and how these skills can be fostered.

## 3.2 The role of different competences in the development of growth trajectories

Manuscript 2 explored how the separate analysis of different competences can contribute to the explanation of growth trajectory patterns that had been found in previous research. The study revealed that growth trajectories in first-grade students may be more diverse than studies suggested which were placed in homogeneous settings (Aunola, Leskinen, Lerkkanen, & Nurmi, 2004) or which used normative cutoff scores for the categorization of developmental groups (Bodovski & Farkas, 2007; Morgan et al., 2009). These studies found increasing or at least stable differences in student scores over the course of several years. In contrast to these findings, the data-driven analyses of precursor and curricular competences within the first grade provides strong evidence for catch-up effects of students in all competences, which had in part also been found by Jordan et al. (2007, 2006).

If the results are replicated in other contexts, the study has wide-ranging implications both for research and practice. The results support the assumption of a sequential, gateway-like development of mathematics skills. Therefore, it seems apparent that students with weaknesses in precursor competences need fostering in this area before they can attain a comprehension of sophisticated mathematical concepts. Nevertheless, a noteworthy proportion of students with low initial scores managed to catch up to their continuously high-performing peers, which demonstrates that favorable performance outcomes are also reachable for students with little prior knowledge. At the same time, students who were proficient in precursor competences and had average Computation skills at the beginning of first grade then divided into trajectory groups with very different learning growth in the computation domain. Monitoring their progress over time is advisable. Lastly, students with largely above-average initial scores in a competence were also very likely to have very high scores at the end of the school year, and students with high initial precursor scores were very unlikely to belong to a Computation class with unfavorable outcome.

Important limitations of the approach used in this study should be considered which relate to the issues that were described for slope estimates. First, although the obtained latent class models fit the data well, the trajectory classes still describe aggregated data: Students were assigned to the classes that *best* fit their individual development of scores of all obtained classes, but not all student data fit these classes well. Second, trajectory group characteristics and students' group memberships were obtained post hoc (after all data from the study had been collected). In application scenarios, one of the defining features of progress monitoring is the opportunity to react more quickly to students' learning processes than in traditional (static) assessment scenarios. If knowledge about typical learning trajectories is to be feasible for educators in progress monitoring scenarios, ways of reliably classifying students within short time frames need to be found. Finally, the results of the study only apply to first grade, and it remains unknown whether similar catch-up patterns can still be observed in grade two and higher. Longer-term longitudinal studies suggest that trajectory paths grow increasingly stable in higher grades (Aunola et al., 2004; Geary, Hoard, Nugent, & Bailey, 2012; Jordan et al., 2007; Morgan et al., 2009).
The study nonetheless implies that researchers should analyze growth trajectories in mathematics from a more differentiated stance than has mostly been the case to date. Using data-driven methods to categorize students into trajectory groups avoids the danger of falsely categorizing students with meaningfully different learning growth (but similar initial competence levels) into one group. Furthermore, analyzing diverse competences in addition to overall scores has strong potential to advance the research base in mathematics developmental models.

# 4 Conclusions

In summary, this dissertation provides evidence for the newly-created test concepts as means of monitoring students' mathematics progress in first and second grade. Assessing both precursor and grade-level curricular competences, the tests provide educators with valid and reliable diagnostic information about students' performance and their development. Implementation of the tests, with short assessment times, has been shown to be feasible in general-education settings, and teachers reported to use the results for diverse purposes. Finally, the dissertation extends the research on mathematics growth trajectories by identifying preconditions for diverse trajectory paths in first grade. Further research is required to explore possibilities and limitations of using the progress monitoring results for instructional adjustments and, thus, enhanced student learning.

# 5

# References

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. doi:10.1016/j.jsp.2012.09.004

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

Aunola, K., Leskinen, E., & Nurmi, J.-E. (2006). Developmental dynamics between mathematical performance, task motivation, and teachers' goals during the transition to primary school. *British Journal of Educational Psychology*, *76*(1), 21–40. doi:10.1348/000709905X51608

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

Bast, J., & Reitsma, P. (1997). Matthew effects in reading: A comparison of latent growth curve models and simplex models with structured means. *Multivariate Behavioral Research*, *32*(2), 135–167. doi:10.1207/s15327906mbr3202_3

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

Bodovski, K., & Farkas, G. (2007). Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction. *The Elementary School Journal*, *108*(2), 115–130. doi:10.1086/525550

Bryant, D. P., Bryant, B. R., Gersten, R., Scammacca, N., & Chavez, M. M. (2008a). Mathematics intervention for first- and second-grade students with mathematics difficulties: The effects of Tier 2 intervention delivered as booster lessons. *Remedial and Special Education*, *29*(1), 20–32. doi:10.1177/0741932507309712

Bryant, D. P., Bryant, B. R., Gersten, R., Scammacca, N., & Chavez, M. M. (2008b). Errata to: Mathematics intervention for first- and second-grade students with mathematics difficulties: The effects of Tier 2 intervention delivered as booster lessons. *Remedial and Special Education*, *29*(4), 252–252. doi:10.1177/0741932508318665

Bus, A. G., & van IJzendoorn, M. H. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology*, *91*(3), 403–414. doi:10.1037/0022-0663.91.3.403

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

Christ, T. J., Scullin, S., Tolbize, A., & Jiban, C. L. (2008). Implications of recent research: Curriculum-based measurement of math computation. *Assessment for Effective Intervention*, *33*(4), 198–205. doi:10.1177/1534508407313480

Christ, T. J., Zopluoglu, C., Monaghen, B. D., & Van Norman, E. R. (2013). Curriculum-based measurement of oral reading: multi-study evaluation of schedule, duration, and dataset quality on progress monitoring outcomes. *Journal of school psychology*, *51*(1), 19–57. doi:10.1016/j.jsp.2012.11.001

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.

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.

Dehaene, S. (2001). Précis of the number sense. *Mind & Language*, *16*(1), 16–36. doi:10.1111/1468-0017.00154

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. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional children*, *52*(3), 219–232.

Dyson, N. I., Jordan, N. C., & Glutting, J. (2013). A number sense intervention for low-income kindergartners at risk for mathematics difficulties. *Journal of learning disabilities*, *46*(2), 166–81. doi:10.1177/0022219411410233

Ehmke, T., Blum, W., Neubrand, M., Jordan, A., & Ulfig, F. (2003). Wie verändert sich die mathematische Kompetenz von der neunten zur zehnten Klassenstufe? [How does mathematical competence change between grade nine and ten?]. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, R. Neubrand, … U. Schiefele (Eds.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres [PISA 2003. Studies on the development of competence in the course of a school year]* (pp. 63–85). Münster: Waxmann.

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students’ reading with differing achievement levels. *Journal of Educational Research*, *102*(6), 453–462. doi:10.3200/JOER.102.6.453-462

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

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, *90*(1), 37–55. doi:10.1037/0022-0663.90.1.37

Förster, N. (2013). *Lernverlaufsdiagnostik in der Grundschule – Konstruktion und Evaluation eines Verfahrens zur Dokumentation von Lernverläufen im Lesen [Learning progress assessment in primary school – construction and evaluation of an approach to document learning progress in reading]*. University of Münster, Germany.

Freebody, P., & Tirre, W. C. (1985). Achievement outcomes of two reading programmes: An instance of aptitude-treatment interaction. *British Journal of Educational Psychology*, 55(1), 53–60. doi:10.1111/j.2044-8279.1985.tb02606.x

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

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2012). The early prevention of mathematics difficulty: its power and limitations. *Journal of learning disabilities*, *45*(3), 257–69. doi:10.1177/0022219412442167

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2012). Mathematical cognition deficits in children with learning disabilities and persistent low achievement: A five-year prospective study. *Journal of Educational Psychology*, *104*(1), 206–223. doi:10.1037/a0025398

Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children*, *78*(4), 423–445.

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.

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

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*(3), 297–313. doi:10.3102/00346543059003297

Jordan, N. C., Kaplan, D., Locuniak, M. N., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice*, *22*(1), 36–46. doi:10.1111/j.1540-5826.2007.00229.x

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. R. (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

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., & 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

Lee, Y.-S., Lembke, E. S., Moore, D., Ginsburg, H. P., & Pappas, S. (2012). Item-Level and Construct Evaluation of Early Numeracy Curriculum-Based Measures. *Assessment for Effective Intervention*, *37*(2), 107–117. doi:10.1177/1534508411431255

Lehrplan Mathematik für die Grundschulen des Landes Nordrhein-Westfalen [Mathematics curriculum for primary schools in the federal state of North Rhine-Westphalia]. (2008). Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. Retrieved from http://www.standardsicherung.schulministerium.nrw.de/lehrplaene/upload/klp_gs/GS_LP_M.pdf

Leppänen, U., Niemi, P., Aunola, K., & Nurmi, J.-E. (2004). Development of reading skills among preschool and primary school pupils. *Reading Research Quarterly*, *39*(1), 72–93. doi:10.1598/RRQ.39.1.5

Lerkkanen, M.-K., Rasku-Puttonen, H., Aunola, K., & Nurmi, J.-E. (2004). Reading performance and its developmental trajectories during the first and the second grade. *Learning and Instruction*, *14*(2), 111–130. doi:10.1016/j.learninstruc.2004.01.006

Lorenz, C., & Artelt, C. (2009). Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik [Domain specificity and stability of diagnostic competence among primary school teachers in the school subjects of german and mathematic. *Zeitschrift für Pädagogische Psychologie*, *23*(3), 211–222. doi:10.1024/1010-0652.23.34.211

*Math standards and expectations*. (2012). Retrieved from http://www.nctm.org/standards/content.aspx?id=314

McDonald Connor, C., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., … Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child × instruction interactions on first graders' literacy development. *Child Development*, *80*(1), 77–100.

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

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities*, *42*(4), 306–321. doi:10.1177/0022219408331037

Morgan, P. L., Farkas, G., & Wu, Q. (2011). Kindergarten children's growth trajectories in reading and mathematics: Who falls increasingly behind? *Journal of Learning Disabilities*, *44*(5), 472–488. doi:10.1177/0022219411414010

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J.-E., & Kirby, J. R. (2005). Development of individual differences in reading: Results from longitudinal studies in english and finnish. *Journal of Educational Psychology*. American Psychological Association. doi:10.1037/0022-0663.97.3.299

Polignano, J. C., & Hojnoski, R. L. (2012). Preliminary evidence of the technical adequacy of additional curriculum-based measures for preschool mathematics. *Assessment for Effective Intervention*, *37*(2), 70–83. doi:10.1177/1534508411430323

Salaschek, M., & Souvignier, E. (under review). Web-based mathematics progress monitoring in second grade. *Journal of Psychoeducational Assessment.*

Salaschek, M., & Souvignier, E. (accepted). Web-based progress monitoring in first grade mathematics. *Frontline Learning Research*.

Salaschek, M., Zeuch, N., & Souvignier, E. (under review). Mathematics growth trajectories in first grade: Cumulative vs. compensatory patterns and the role of number sense. *Learning and Individual Differences.*

Scarborough, H. S. (2009). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In F. Fletcher-Campbell, J. Soler, & G. Reid (Eds.), *Approaching difficulties in literacy development: Assessment, pedagogy and programmes* (p. 23). SAGE Publications Inc.

Schulgesetz für das Land Nordrhein-Westfalen [Education law for the federal state of North Rhine-Westphalia] (2013). Retrieved from http://www.schulministerium.nrw.de/docs/Recht/Schulrecht/Schulgesetz/Schulgesetz.pdf

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

Sood, S., & Jitendra, A. K. (2013). An exploratory study of a number sense program to develop kindergarten students' number proficiency. *Journal of learning disabilities, 46*(4), 328–46. doi:10.1177/0022219411422380

Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research & Practice, 15*(3), 128–134. doi:10.1207/SLDRP1503_2

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

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology, 38*(6), 934–947. doi:10.1037/0012-1649.38.6.934

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. doi:10.1026/0049-8637/a000011

# PART II

## Manuscript 1

## Web-based progress monitoring in first grade mathematics

# Web-based progress monitoring in first grade mathematics

Martin Salaschek & Elmar Souvignier

University of Münster, Germany

Fliednerstrasse 21, 48149 Münster, Germany, martin.salaschek@uni-muenster.de

## 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.

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.

| | 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.

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.

*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.

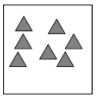
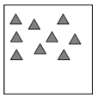
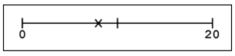
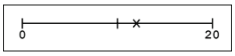
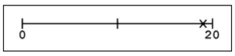
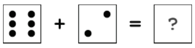
Because of student mobility or sick absentees, some data were missing (progress monitoring tests: 0%-11%, $M_{missing}$ = 1.8%; paper pencil tests: 0%-3.6%, $M_{missing}$ = 1.7%; teacher ratings: 4.5%-23.2%, $M_{missing}$ = 12.6%). We used multiple imputation with five imputed data sets to handle missing 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).

## 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.

*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"* |  <br> <br> | 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.

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.

## 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.

[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.

# 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).

*Table 3.* Internal consistencies of progress monitoring overall scores and competence scores

| | α | | | |
|---|---|---|---|---|
| progress monitoring | 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 ≤ $r$ ≤ .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 ≤ $r$ ≤ .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 ≤ $r$ ≤ .68. Later progress monitoring tests related to the pp2 and pp3 scores to a somewhat lesser degree (see Table 4).

*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

[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.

between the progress monitoring scores and ratings provided at the beginning of grade 1 (teacher rating 1; .29 ≤ *r* ≤ .42). Correlations with ratings provided at the end of grade 1 were substantially higher (teacher rating 2; .54 ≤ *r* ≤ .64) and remained stable for ratings provided at the end of grade 2 (teacher rating 3; .54 ≤ *r* ≤ .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 | teacher ratings | | |
|---|---|---|---|
| monitoring | 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 scale from 1, *strongly disagree* to 5, *strongly agree*). 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 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

| | overall score | | Basic Precursors | | Advanced Precursors | | Computation | |
|---|---|---|---|---|---|---|---|---|
| progress monitoring | *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).

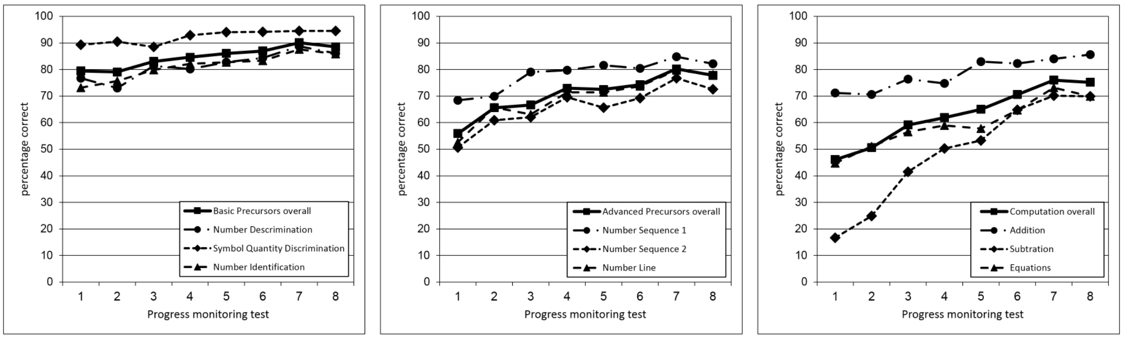


*Figure 2.* Growth rates for single measures in study 2 (*n* = 153).

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 paper pencil scores 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 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

# 6

# 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.

## Appendix

Descriptive statistics and growth rates for competences, study 1

| | overall score | | Basic Precursors | | Advanced Precursors | | Computation | |
|---|---|---|---|---|---|---|---|---|
| progress monitoring | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| time 1 | 61.9 | 14.4 | 74.7 | 14.4 | 54.8 | 20.8 | 52.7 | 19.5 |
| time 2 | 64.8 | 14.5 | 77.2 | 13.0 | 63.3 | 21.9 | 49.9 | 19.2 |
| time 3 | 66.1 | 14.3 | 80.7 | 12.9 | 63.5 | 21.5 | 49.7 | 19.1 |
| time 4 | 69.2 | 14.6 | 80.4 | 12.2 | 68.0 | 22.2 | 55.7 | 21.5 |
| time 5 | 69.4 | 15.3 | 79.9 | 13.5 | 67.2 | 23.2 | 57.8 | 21.0 |
| time 6 | 70.3 | 15.0 | 78.4 | 13.9 | 68.7 | 21.5 | 61.0 | 21.3 |
| time 7 | 70.8 | 15.8 | 81.5 | 13.8 | 68.1 | 22.4 | 59.5 | 22.7 |
| time 8 | 73.1 | 15.8 | 81.4 | 13.4 | 72.0 | 23.0 | 63.1 | 23.0 |
| Growth rate | 0.7 | | 0.4 | | 0.9 | | 0.9 | |

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

# PART III

## Manuscript 2

## Mathematics growth trajectories in first grade: Cumulative vs. compensatory patterns and the role of number sense

# Mathematics growth trajectories in first grade: Cumulative vs. compensatory patterns and the role of number sense

Martin Salaschek[a], Nina Zeuch[a], Elmar Souvignier[a]

[a] University of Münster, Fliednerstrasse 21, 48149 Muenster, Germany

Corresponding author: Martin Salaschek
salaschek@uni-muenster.de
+49-251-8334300

Author Note
The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors received no financial support for the research, authorship, and/or publication of this article.

# Abstract

We examined mathematics growth trajectories in first grade for overall achievement and three separate competences (Basic Precursors, Advanced Precursors, Computation). 153 German students computed seven web-based progress monitoring tests during the school year. Latent class growth analysis (LCGA) provided evidence for mainly cumulative patterns of performance development: In all competences, we found groups of initially high-performing students with the highest end scores and groups of initially low-performing students with little or no growth. In addition, compensatory patterns with groups of initially lower-performing students and steep growth were found. For precursor competences, these catch-up groups did not have increased odds of belonging to low-end outcome groups in higher competences. Given that students with similar initial performance differed substantially in their learning growth, monitoring students' progress in educational settings in narrow intervals of time seems commendable.

*Keywords:* growth trajectories; mathematics; number sense; progress monitoring

# 1 Introduction

During the last decade, there has been growing interest in the developmental dynamics of children's math performance during the first years of education. Several research groups have used longitudinal data to model learning trajectories of different groups of students, providing remarkable new insight into the stability of students' long-term mathematical performance (e.g., Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Bodovski & Farkas, 2007; Geary, Hoard, Nugent, & Bailey, 2012; Jordan, Kaplan, Oláh, & Locuniak, 2006; Morgan, Farkas, & Wu, 2009). All of these sophisticated studies found that students with low math performance at the beginning of the study were likely to show substantially less learning growth than their peers with higher initial performance. However, most of the studies used measures of overall mathematical performance to categorize students into trajectory groups over several years. Less is known about the shorter-term heterogeneity in the development of early math skills and about the interplay of different competences. The aim of the present study thus was to examine different growth trajectories of students' overall math performance and of several separate competences in grade 1.

## 1.1 Early mathematics skills

Basic mathematical competences, such as knowledge about quantities and numbers, counting abilities, or basic arithmetic facts, start to develop before school entry and broaden during the first years of formal education. These competences, commonly referred to as *precursor* or *number sense* competences (for an overview, see Berch, 2005), have not been consistently defined, but there is widespread agreement on their importance for students' further mathematical development (e.g., Chard et al., 2005; Kolkman, Kroesbergen, & Leseman, 2013; Krajewski & Schneider, 2009; Locuniak & Jordan, 2008; Missall, Mercer, Martínez, & Casebeer, 2012). Number sense sub-skills can be categorized with regard to the sequence of their development. First, children learn several basic skills, such as discriminating between quantities and identifying numbers as quantities. These skills usually develop between the age of 2 and 5 and can be acquired independently from each other (Krajewski & Schneider, 2009). We refer to these skills as *Basic Precursors*. Basic precursors make way for more advanced skills, e.g., recognizing number patterns or identifying a number on a number line (Dehaene, Piazza, Pinel, & Cohen, 2003; Siegler & Booth, 2004). Such tasks require the integration of several basic quantity-related skills and should mainly be developed before school entrance, as they are highly predictive of longer-term math achievement (Krajewski & Schneider, 2009). We refer to these skills as *Advanced Precursors*. Finally, developing *Computation* competence, such as addition and subtraction skills, is the main curricular goal in math instruction in the first school years. Fundamental understanding of addition and subtraction forms the basis for the most important curricular goals of the elementary school grades.

However, such skill developments are not strongly sequential. As Krajewski and Schneider (2009) note, children can reach different competence levels for smaller and larger numbers, making it hard to accurately determine a child's competence level. But little is known about typical sequences of number sense development because children learn fast, and longitudinal studies assessing several competences in short intervals are rare. Knowledge about varying growth trajectories in different math competences may help researchers and educators better understand the preconditions of a favorable or unfavorable long-term development, which can in turn be used to improve systematic early interventions.

## 1.2 Growth trajectories in mathematics

Growth trajectories in general-education mathematics have been more intensely studied using longitudinal multivariate methods for about a decade (Aunola et al., 2004; Bodovski & Farkas, 2007; Geary et al., 2012; Jordan et al., 2006; Morgan et al., 2009). Main aim of these studies is to describe how students systematically differ in their skills development over time and what characterizes different trajectory groups. This knowledge may help with early identification of students at risk of developing persistent math difficulties (MD) and with designing trajectory-specific interventions.

### 1.2.1 Growth trajectories of overall math performance

For the description of growth trajectories among students, cumulative or compensatory developmental patterns are discussed in the literature. Cumulative patterns, i.e., increasing differences in performance and variance between students over time, have been described by a number of studies in reading and mathematics (e.g., Bast & Reitsma, 1997; Bodovski & Farkas, 2007; Kempe, Eriksson-Gustavsson, & Samuelsson, 2011; Leppänen, Niemi, Aunola, & Nurmi, 2004; Williamson, Appelbaum, & Epanchin, 1991). This effect, with biblical reference, is also called *Matthew effect* (Stanovich, 1986). In contrast, a compensatory effect describes a pattern where initially less skilled students, with the beginning of formal instruction, show higher growth rates than initially higher-skilled children. As a consequence, the achievement gap between students narrows over time. This effect has been described in a number of studies for reading (e.g., Aarnoutse & van Leeuwe, 2000; Parrila, Aunola, Leskinen, Nurmi, & Kirby, 2005; Phillips, Norris, Osmond, & Maynard, 2002), but evidence for mathematics is scarce. These research findings with mixed evidence may suggest that there are certain conditions (e.g., certain characteristics of schooling or certain patterns of precursor abilities) under which students follow either trajectory path.

The number of studies that use person-oriented variables to identify distinct mathematics growth trajectories increases slowly. Aunola et al. (2004) followed 194 Finnish children from kindergarten to the end of grade 2. Using growth mixture modeling, two trajectory groups were obtained: A *high performers* class was defined by a high overall level of performance and a high and positive growth rate. A *low performers* class was characterized by lower overall performance and a lower (but also positive) growth rate. Therefore, the results of the study suggest a cumulative pattern of math performance. No distinct small group that performs well below average was found in this study, as would be suggested by studies that identify groups of 5-20% of students with specific or broad learning difficulties (see Geary, 2004 and Mazzocco, 2005 for discussions of prevalences). This may be due to the homogeneous characteristics of the study sample.

Jordan and colleagues (2007, 2006) assessed the performance of about 400 students from kindergarten to the middle of grade 1 in a more heterogeneous sample. Using growth mixture modeling, Jordan et al. (2007) found three classes of overall number sense performance that were named by their outcome level and slope characteristics: a *low/flat* group displayed low performance at the last time point and only small competence gains over time; a *middle/steep* class started only slightly higher than the first group but displayed significantly steeper growth; finally, a *high/flat* class started out in kindergarten scoring about twice as high as the first group and displayed growth levels in-between the other two groups (scoring highest at the end of the study). Although this pattern may suggest a compensatory effect for some children, there were ceiling effects for the highest-scoring class, and it is unknown whether this class would have shown higher growth if more complex tasks had been used. In their 2006 study, using only the four time points in kindergarten, Jordan and colleagues found three very similar groups. However, growth of the *high* group was slightly higher than growth of the *middle* group during this shorter time frame.

In contrast to these studies that use data-driven methods to obtain trajectory characteristics and group sizes, several studies used normative performance criteria to categorize students into trajectory groups. Morgan et al. (2009) analyzed performance level and growth from kindergarten to fifth grade for children showing different MD patterns in kindergarten. With extensive data from the US-representative Early Childhood Longitudinal Study-Kindergarten Cohort study (ECLS-K), the lowest-performing 10% of students were categorized as having MD in fall and/or spring of kindergarten. Students who did not display MD in either fall or spring of kindergarten showed the highest performance level and growth, followed by students displaying MD in fall kindergarten only, students displaying MD in spring kindergarten only, and finally students displaying MD in both fall and spring of kindergarten.

Bodovski and Farkas (2007) also used ECLS-K data to analyze students' performance from kindergarten to third grade. The authors divided students into four equally-sized groups according to their kindergarten fall performance. This approach yielded in very different categorizations than the approach used by Morgan and colleagues (2009) because the no-MD group in the Morgan et al. study roughly comprised the proportion of students who were represented in the three higher quartiles in the Bodovski and Farkas study. Vice versa, all three MD groups of the Morgan et al. (2009) study were roughly represented in just the lowest quartile in the Bodovski and Farkas (2007) study. Nonetheless, major conclusions about growth trajectories were similar: the higher the initial performance level, the higher the growth – with the exception of the two highest quartiles, which did not differ in their growth.

In sum, all five studies found at least one low-performing group of students with low learning growth over time, and one or more groups with higher overall performance and higher learning growth. These results seem to indicate that students generally follow a cumulative pattern during early development in mathematics.

### 1.2.2 Growth trajectories of discrete math competences

Growth trajectories of discrete skills were analyzed in detail by Geary et al. (2012), who followed 177 students from first through fifth grade and assessed several measures of math achievement, along with other competences. With respect to their mathematics achievement, the authors categorized students as *typically achieving* (TA), persistently *low achieving* (LA), or having a *mathematics learning disability* (MLD). In a precursor *number sets* task, where children were asked to identify sets of symbols and Arabic digits that add up to a certain quantity, the LA and MLD groups reduced the performance gap from first to second grade, but group differences were constant after that. In a number line task, the gap between all groups narrowed over several years, and the difference between the LA and TA group was not significant anymore in fifth grade. In an addition task, strategy use was recorded, and results were separately reported for simpler (procedural) and more advanced (decomposition and retrieval) strategies. For procedural strategy use (e.g., counting fingers), the performance gap had mainly closed at second grade. For advanced strategy use, the gap showed a typical fan-spread pattern from first to second grade.

Jordan et al. (2006) also briefly reported results from three of their number sense measures. In some of the measures of the three-class solutions, the highest rates of growth could be observed among a group of students which started lower than the highest overall-performing trajectory class.

Results from both studies suggest that there may be distinct growth trajectories for different tasks or competences. For precursor skills, some compensatory patterns were observed, whereas mainly cumulative patterns were found for more advanced competences.

### 1.2.3 Stability of growth trajectories over time

The overall time spans and assessment intervals of the presented studies vary from four time points in one kindergarten year (Jordan et al., 2006) to 5 time points in about six years (Morgan et al., 2009), so conclusions about the developmental dynamics differ. Geary et al. (2012) found large performance changes between the groups in the first year of the study and much smaller changes after that. Results from Morgan et al. (2009) also suggest greater developmental dynamics when children are younger. In their study, about half of the children categorized as having MD in the fall of kindergarten did not fall in the MD categorization six months later. This group of less-persistent MD then showed significantly higher learning growth throughout the years than their peers with MD at both fall and spring of kindergarten. It therefore seems commendable to inspect early developments more closely.

### 1.2.4 Trajectory classifications methods

The presented studies differ in the way that students were categorized into trajectory groups. Most studies used cutoff scores to categorize students into trajectory groups. Among the benefits of this normative approach is the possibility of classifying students' performance after a single assessment (e.g., *risk* vs. *no-risk*). A priori risk estimations can then be evaluated concerning specificity and sensitivity at a later time point.

Of the presented studies, only Aunola et al. (2004) and Jordan et al. (2007, 2006) used latent growth modeling for categorizations that best fit the data. Data-driven classification methods in longitudinal settings allow more exact estimations of the proportion of students that follow a specific trajectory path. In the studies by Aunola et al. (2004) and Jordan et al. (2007), about 35% of the students belonged to a homogeneous low-performance group. Normative classifications in these data might have produced biased or less clear results. Moreover, as Morgan et al. (2009) and Geary et al. (2012) point out, students with similarly low performance at an early assessment show diverse subsequent learning growth. Thus, observing the persistence of difficulties in math seems to be the deciding factor in risk estimations, and latent growth modeling provides suitable means to handle this concern.

## 1.3 Aims of the study

Several recent studies examined the development of math performance in primary education. Analyses in these studies revealed mainly cumulative growth patterns, but results differed with regard to more specific trajectory characteristics. Several studies showed catch-up effects for some students, particularly when sub-skills were assessed.

In our study we focused on learning growth trajectories for different competences in first grade. We took a closer look at the beginning of formal education because we expected higher variability in performance level and growth than in higher grades. Moreover, growth trajectory analyses at the transition from informal to formal schooling may provide more insight about the importance of different precursor skills for the development of curricular skills. Thus, we used a progress monitoring tool which assesses skill development separately for diverse competences (Basic Precursors, Advanced Precursors, and Computation, in addition to total scores). Based on the previous research in this area, we pursued three research questions.

First, we used latent growth curve models (LGCM; e.g., Bollen & Curran, 2006) to determine if growth in first grade differs significantly between individuals. Based on the findings by Aunola et al. (2004) and Jordan et al. (2007, 2006), we expected to find significant variance in the slopes of students' overall scores.

Second, latent class growth analysis (LCGA; Jung & Wickrama, 2008) was performed to obtain data-derived trajectory groups. We expected to find mostly cumulative growth patterns, i.e., trajectory groups with higher starting performance should display steeper slopes than trajectory groups with lower starting performance. With regard to research by Jordan et al. (2007), we also expected to find students with average or below-average overall starting performance and steep slopes (‘catch-up groups’).

Third, we analyzed whether related trajectory groups could also be found for specific competences. We expected Precursor trajectories to be more uniform than Computation trajectories because Precursor competences were expected to be mainly developed by the start of formal schooling.

Fourth, given that we analyzed the competences separately, we were interested in the stability of trajectory group classifications, i.e., if students belonged to similarly-characterized groups across the competences. Assuming a gateway role of Precursor skills, we expected that the odds of belonging to a high-performing Computation group would be significantly decreased for students with persistently low Precursor skills.

# 2

# Method

## 2.1 Participants and procedure

A total of 153 first-grade students (46% girls, *M* = 6.72 years at first assessment) participated in the study and completed short math tests every three weeks from November 2011 to May 2012. The study took place in rural and urban areas of Germany. 9% of the students were categorized as having a migration background by their teachers, defined as speaking another language than German at home.

Students were originally examined eight times. Test scores decreased from test 7 to the last test, however (Authors, in press), a result that was also observed by other researchers when testing frequently (Förster & Souvignier, 2011; Hampton et al., 2012). Given that overall test-to-test increases were observed for all other tests, we assumed that the decrease in the last test was due to motivational processes and excluded test 8 from all further analyses.

## 2.2 Measures of mathematical performance

We aimed to examine the performance level and growth of math performance across several time points in one school year. Thus, an assessment tool was necessary which reliably assesses relevant competences and performance growth over time. We developed a web-based progress monitoring tool for this purpose, following the theoretical framework of curriculum-based measurement (CBM; Espin, Mcmaster, Rose, & Wayman, 2012; Fuchs & Fuchs, 1998). The test concept comprised nine types of tasks (*measures*) in three competences. The different measures in the test were designed to assess Basic

Precursors, Advanced Precursors, and Computation competences as described in the Introduction.

Four parallel tests (A, B, C, D) were created and conducted in the sequence A-D, A-C. Students worked on the computer-based tests independently during regular classroom hours. The test format was specifically designed for first-grade students, and all necessary instructions were provided via headphones. The tests were self-paced, and median assessment times ranged from 15.58 minutes (*SD* = 4.67) in test 1 to 10.11 minutes (*SD* = 3.13) in test 7. The level of difficulty was aligned to the curricular goals, with most tasks including quantities from 1-20. Thus, the percentage of correct answers could be used to estimate proficiency in each competence.

Adjacent-test retest reliability of the parallel test forms has been demonstrated to be adequate for three-week test intervals ($r_M$ = .78; Authors, in press), and end-of-year validity of the single tests was high ($r_M$ = .63 with the standardized school achievement test *DEMAT 1+*, Krajewski, Küspert, Schneider, Deimann, & Kastner-Koller, 2002). Internal consistencies of the single tests were narrow-ranging and adequate for overall test scores ($\alpha_M$ = .86). Internal consistencies for the single competences were lower, but still at an acceptable level (Basic Precursors: $\alpha_M$ = .63; Advanced Precursors: $\alpha_M$ = .80; Computation: $\alpha_M$ = .74).

Basic Precursors (total possible score: 20) included the measures Number Discrimination, Symbol Quantity Discrimination, and Number Identification. In Number Discrimination, two numbers were displayed on the computer screen, and children were asked to select the larger one. In Symbol Quantity Discrimination, two pictures with different amounts of shapes were displayed, and children were asked to select the picture with more shapes. In Number Identification, a spoken number was presented via headphones, and five numbers were displayed on the screen. Children were asked to select the number from the audio. Skills needed for these tasks should mainly develop before school entrance, and students should quickly be proficient in them early in the school year.

Advanced Precursors (total possible score: 17) included the measures Number Sequence 1, Number Sequence 2, and Number Line. In Number Sequence 1, two ascending or descending numbers in steps of one were presented, and children were asked to count onwards and select the correct next number from four alternatives. Number Sequence 2 was the same task, but with numbers ascending or descending in steps of two. In Number Line, a spoken number between 1 and 20 was presented via headphones, and three number lines (starting with 1, ending with 20) were displayed on the screen. The number lines had marks at different numerical positions, and children were asked to select the number line marking the number that was given via audio.

Computation (total possible score: 15) included the measures Addition, Subtraction, and Equation. The former two measures consisted of regular addition and subtraction problems in the range of 1 to 20, and the correct solution had to be selected from four alternatives. In Equation, an addition problem was presented with dice (e.g., one die showing two dots, a plus sign, and one die showing three dots), along with three arithmetic problems with numbers. Children were asked to select the analogous problem with the same solution (e.g., *4 + 1*). Computation skills with Arabic notation are not expected to be developed before school entrance but are central to the math curriculum in first grade, and students should be proficient in these tasks by the end of the school year.

# 3

# Results

Results are presented in two steps. To investigate general growth trajectories for overall scores and each competence, we used latent growth curve models (LGCM) to obtain intercepts (individual initial status) and slopes (individual growth parameters) for the whole sample. We then performed latent class growth analyses (LCGA) on overall scores and each competence to ascertain whether there are distinguishable latent classes of students who show different initial statuses and growth trajectories. All growth curve analyses were conducted within a multilevel framework using Mplus 6.1 (Muthén & Muthén, 1998-2010), taking into account school class membership of students in order to obtain correct significance test results (see Cohen, Cohen, West, & Aiken, 2003).

Analysis of missing data revealed that no more than 11 percent of test data were missing at any time point. Missing data was handled by means of full information maximum likelihood (FIML), which uses all data points available for each occasion and does not impute any data (Enders, 2001; Graham, 2009; Schlomer, Bauman, & Card, 2010).

## 3.1 Latent growth models

LGCM were computed to investigate general growth of overall scores and the three competences. Fit indices are reported in Table 1. Estimates of means and variances for intercepts and slopes are reported in Table 2. Fit can be regarded as good to satisfying for all models (for interpretation of fit indices see Hu & Bentler, 1999 and Schermelleh-Engel, Moosbrugger, & Müller, 2003). All intercepts and slopes were significantly higher than zero which indicates considerable growth across time. Additionally, for Basic Precursors, the correlation of intercept with slope was significantly lower than zero, which means that students who showed high initial performance had less steep growth over time than students who displayed lower initial performance. This pattern seems to be due to ceiling effects for Basic Precursors. Thus, the negative correlation of intercept with slope seems to be artificial and should not be further interpreted.

Inclusion of quadratic parameters revealed no considerable improvement in model fit, and inspection of graphs from LGCMs suggested no systematic non-linear growth. Thus, linear growth was considered appropriate. Variances of slopes being significantly higher than zero (except for Basic Precursors, where the slope was marginally significant) suggested that there was relevant inter-individual variation within slopes. Consequently, analyzing the data with regard to distinguishable subgroups of students with different slopes seemed indicated, and we performed LCGA.

*Table 1.* Fit indices from LGCM for overall scores and competences

| | $\chi^2$ (df, N) | CFI | TLI | RMSEA |
|---|---|---|---|---|
| Overall Scores | 45.25 (22, 153)** | 0.99 | 0.99 | 0.08 |
| Basic Precursors | 28.95 (23, 153) | 0.99 | 0.99 | 0.04 |
| Advanced Precursors | 53.00 (23, 153)** | 0.97 | 0.97 | 0.09 |
| Computation | 35.44 (23, 153)* | 0.99 | 0.99 | 0.06 |

*Note.* CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation. * $p < .05$. ** $p < .01$.

*Table 2.* Estimates from LGCM for intercepts, slopes and correlation of intercepts with slopes

| | i (SE) | var i (SE) | s (SE) | var s (SE) | i × s (SE) |
|---|---|---|---|---|---|
| Overall Scores | 32.69 (0.72)** | 33.95 (2.88)** | 1.71 (0.19)** | 0.51 (0.15)** | -0.13 (0.16) |
| Basic Precursors | 15.75 (0.24)** | 3.67 (0.61)** | 0.36 (0.04)** | 0.03 (0.02)+ | -0.66 (0.08)** |
| Advanced Precursors | 10.15 (0.44)** | 7.42 (0.77)** | 0.57 (0.08)** | 0.09 (0.04)* | -0.02 (0.14) |
| Computation | 6.99 (0.16)** | 3.85 (0.74)** | 0.73 (0.11)** | 0.14 (0.04)** | -0.04 (0.11) |

*Note.* i = intercept; var = variance; s = slope. + $p < .10$. * $p < .05$. ** $p < .01$.

## 3.2 Latent class growth analysis

We performed LCGA to explore data-derived growth trajectory groups, which were expected to mainly follow cumulative growth patterns. When conducting LCGA, there are several methods for determining the number of classes. Absolute model fit for LCGA can be judged on the basis of entropy values and average latent class probabilities for most likely latent class memberships (both should be close to 1, cf. Jung & Wickrama, 2008). Relative model fit can be evaluated by Akaike and Bayesian Information Criteria (AIC, BIC,

*Table 3.* Fit indices for Latent Class Growth Analysis

| No. of classes | LL | No. Of free parameters | AIC | BIC | adj. BIC | Entropy | VLMR | adj. VLMR | BLRT |
|---|---|---|---|---|---|---|---|---|---|
| Overall Scores | | | | | | | | | |
| 2 | -3282.89 | 12 | 6589.78 | 6626.14 | 6588.16 | 0.92 | 574.86** | 539.13** | 574.86** |
| 3 | -3184.76 | 15 | 6399.53 | 6444.98 | 6397.51 | 0.91 | 196.25+ | 184.05+ | 196.25** |
| 4 | -3138.91 | 18 | 6313.83 | 6368.38 | 6311.41 | 0.89 | 91.70+ | 86.00+ | 91.70** |
| 5 | -3121.78 | 21 | 6285.56 | 6349.20 | 6282.73 | 0.87 | 34.27* | 32.14* | 34.27** |
| Basic Precursors | | | | | | | | | |
| 2 | -2201.87 | 12 | 4427.74 | 4464.10 | 4426.12 | 0.82 | 311.71** | 292.34** | 311.71** |
| 3 | -2133.74 | 15 | 4297.47 | 4342.93 | 4295.45 | 0.88 | 136.27** | 127.80** | 136.27** |
| 4 | -2130.42 | 18 | 4296.84 | 4351.38 | 4294.41 | 0.79 | 6.63 | 6.22 | 6.63 |
| 5 | -2127.08 | 21 | 4296.15 | 4359.79 | 4293.33 | 0.77 | 6.69 | 6.27 | 6.69 |
| Advanced Precursors | | | | | | | | | |
| 2 | -2623.31 | 12 | 5270.61 | 5306.98 | 5269.00 | 0.91 | 464.01* | 435.18* | 464.01** |
| 3 | -2531.60 | 15 | 5093.20 | 5138.65 | 5091.18 | 0.92 | 183.42 | 172.02+ | 183.42** |
| 4 | -2515.33 | 18 | 5066.66 | 5121.20 | 5064.23 | 0.92 | 32.54 | 30.52+ | 32.54** |
| 5 | -2505.46 | 21 | 5052.92 | 5116.56 | 5050.09 | 0.87 | 19.74 | 18.51 | 19.74** |
| Computation | | | | | | | | | |
| 2 | -2448.89 | 12 | 4921.78 | 4958.15 | 4920.17 | 0.87 | 389.54 | 365.33 | 389.54** |
| 3 | -2374.11 | 15 | 4778.22 | 4823.67 | 4776.20 | 0.90 | 149.57** | 140.27** | 149.57** |
| 4 | -2359.40 | 18 | 4754.80 | 4809.34 | 4752.37 | 0.82 | 29.42 | 27.59 | 29.42** |
| 5 | -2355.22 | 21 | 4752.44 | 4816.08 | 4749.61 | 0.82 | 8.36 | 7.84 | 8.36+ |

*Notes.* LL = Log-likelihood; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; adj. = adjusted; VLMR = Vuong-Lo-Mendell-Rubin-Test; BLRT = Bootstrap Likelihood-Ratio-Test.+ $p < .10$. * $p < .05$. ** $p < .01$.

adjusted BIC) as well as the Vuong-Lo-Mendell-Rubin (VLMR) test and the Bootstrap-Likelihood-Ratio Test (BLRT). VLMR and BLRT compare k-1 and k class solutions. Significant VLMR and BLRT results indicate that the k class solution fits better. Additionally, classes should not be too small (above 1% of the sample size, cf. Jung & Wickrama, 2008). Nylund, Asparouhov, and Muthén (2007) recommend to use BIC and BLRT to decide about the number of latent classes. Moreover, interpretability and meaningfulness of classes and trajectories were drawn on (Jung & Wickrama, 2008; B Muthén & Muthén, 2000; Bengt Muthén, 2003).

*Table 4.* Average latent class probabilities for most likely latent class membership

| | No. of classes | Prob. Class 1 | Prob. Class 2 | Prob. Class 3 | Prob. Class 4 | Prob. Class 5 |
|---|---|---|---|---|---|---|
| Overall Scores | 5 | 0.94 | 0.91 | 0.90 | 0.91 | 0.99 |
| Basic Precursors | 3 | 0.97 | 0.95 | 0.95 | - | - |
| Advanced Precursors | 5 | 0.96 | 0.83 | 0.86 | 0.95 | 0.90 |
| Computation | 5 | 0.96 | 0.85 | 0.85 | 0.80 | 0.90 |

*Notes.* Prob. = Average probability for most likely latent class membership.

Four sets of class models with two through five classes were estimated for overall scores and the three competences (see Table 3 for an overview of the fit indices for all solutions, Table 4 for probabilities of class memberships, and Table 5 for predicted mean scores and slopes for each final class). Inclusion of quadratic parameters revealed no considerable improvement in model fit, and inspection of graphs from LGCM suggested no systematic non-linear growth. Thus, we considered linear growth to be sufficient for all following models. We then explored appropriate class solutions separately for all competences.

### 3.2.1 Overall scores

For overall scores, the five-class solution was selected (see Figure 1 scores and class proportions). We selected this solution because BIC was lower than for four classes, and a significant BLRT was found. Classes were still well interpretable. Average latent class probabilities for most likely latent class membership were well above .90 for all classes and class sizes were above 1%. A six-class solution did not add meaningful trajectories.

*Table 5.* LCGA final solutions: predicted progress monitoring scores at the first and last time point, slopes, and significance of slopes

| | Overall Scores | | | | Basic Precursors | | | | Advanced Precursors | | | | Computation | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | time 1 | time 7 | s | p | time 1 | time 7 | s | p | time 1 | time 7 | s | p | time 1 | time 7 | s | p |
| Class 1 | 79% | 97% | 3.1% | < .001 | 92% | 97% | 0.8% | < .001 | 75% | 94% | 3.0% | < .001 | 62% | 95% | 5.5% | < .001 |
| Class 2 | 67% | 89% | 3.6% | < .001 | 76% | 91% | 2.5% | < .001 | 47% | 87% | 6.7% | < .001 | 40% | 84% | 7.3% | < .001 |
| Class 3 | 54% | 83% | 4.8% | < .001 | 68% | 77% | 1.5% | < .001 | 54% | 68% | 2.4% | .144 | 43% | 63% | 3.3% | < .001 |
| Class 4 | 55% | 66% | 1.9% | .002 | | | | | 37% | 49% | 2.0% | < .001 | 26% | 53% | 4.6% | < .001 |
| Class 5 | 44% | 55% | 2.0% | < .001 | | | | | 27% | 25% | -0.4% | .115 | 36% | 35% | -0.2% | .351 |

*Note.* s = slope.

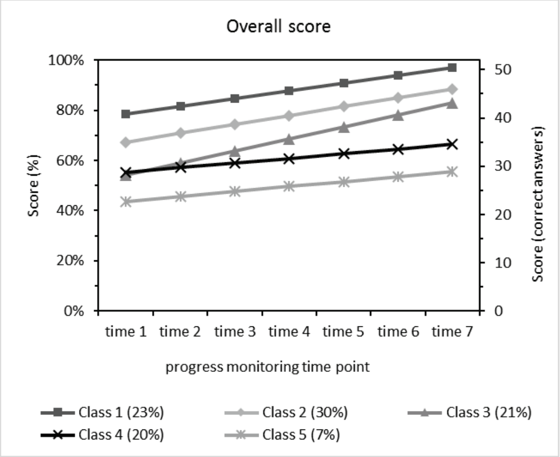


Figure 1. Latent class growth analysis trajectory class scores and proportions for progress monitoring overall scores.

As expected, mostly cumulative overall growth patterns were found. Three classes with about three quarters of the students reached high overall scores by the end of the study. Of these classes, the higher two (*high performers 1* and *2*) displayed strong overall performance with similar growth. Class 3 (*catch-up* class) was characterized by significantly lower starting performance and the steepest slope of all classes. Classes 4 and 5 (*low performers 1* and *2*), with about one quarter of the students, reached considerably lower performance levels and growth rates than all other classes.

### 3.2.2 Basic Precursors

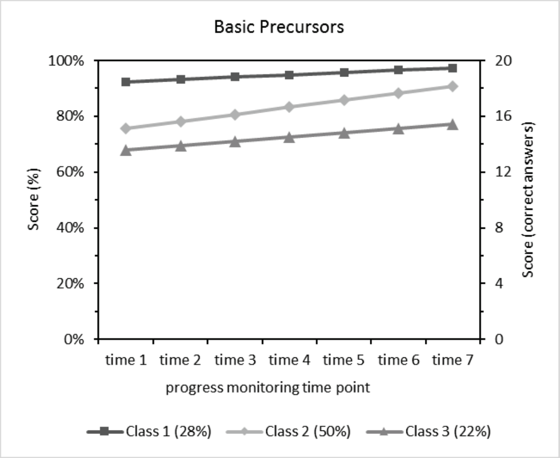


Figure 2. Latent class growth analysis trajectory class scores and proportions for Basic Precursors scores.

For Basic Precursors, the three-class solution was selected (see Figure 2). For this solution, the lowest BIC and a significant BLRT value was found. Classes were well interpretable. Additional classes did not add meaningful trajectories. Average latent class probabilities for most likely latent class membership were close to 1 for all classes, and class sizes were above 1%.

In sum, two classes (class 1, *high performers*; class 2, *catch-up* class) with more than three quarters of the students reached very high Basic Precursors scores at the end of grade 1. While potential growth of class 1 was limited by high initial scores, class 2 started considerably lower and displayed steep growth. Class 3 (*low performers*) started out lowest and displayed low growth throughout the study.

### 3.2.3 Advanced Precursors

For Advanced Precursors, the five-class solution (see Figure 3) was selected. This solution was chosen because BIC was lower than for the four-class solution and a significant BLRT value was found. Classes were still well interpretable. A six-class solution did not add meaningful trajectories. Average latent class probabilities for most likely latent class membership were well above .80 for all classes and class sizes were above 1%.

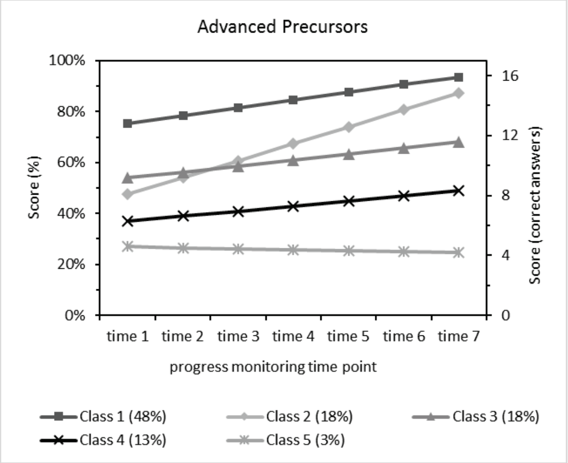


Figure 3. Latent class growth analysis trajectory class scores and proportions for Advanced Precursors scores.

In sum, two classes (class 1, *high performers*; class 2, *catch-up* class) with about two thirds of the students reached high scores by the end of the study. Class 2 was characterized by average starting scores and steep growth, and performance at time point 7 was only slightly lower than for the higher-starting class 1. Class 3 (*low performers 1*), with about one sixth of the students, was characterized by moderate scores at the end of the study. Classes 4 and 5 (*low performers 2* and *3*), with another sixth of the students, reached low scores at time point 7 while displaying moderate to no growth.

### 3.2.4 Computation

In this competence, adjusted BIC of a five-class solution decreased slightly compared to the four-class-solution, and BLRT was marginally significant. Hence, BIC and BLRT suggested a four-class solution, albeit latent class probabilities for most likely latent class membership were well above .80 for four of the five classes and class sizes were above 1%. However, in the four-class solution, classes 4 and 5 (as described below) were categorized as one class (mean scores at t1, t7: 31%, 42%; slope = 1.8%,

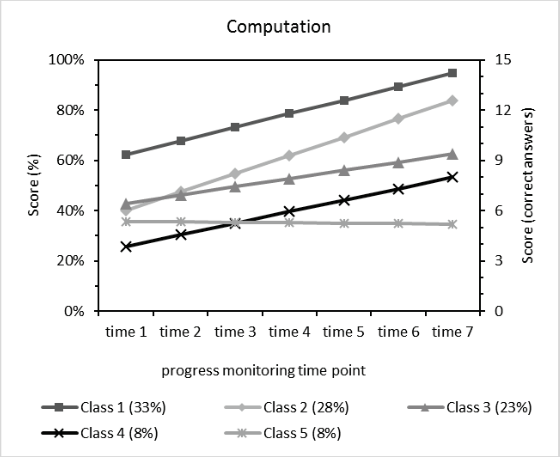


Figure 3. Latent class growth analysis trajectory class scores and proportions for Advanced Precursors scores.

*p* = .282). Given the very distinct growth patterns of classes 4 and 5 in the five-class solution, we selected the five-class solution as suitable (see Figure 4). Characteristics of classes 1 to 3 did not differ noticeably between the two solutions. A six-class solution did not add meaningful trajectories.

In sum, two classes (class 1, *high performers*; class 2, *catch-up* class) which covered 60% of the students reached high scores at the end of the study with steep growth throughout the school year. As was the case with Advanced Precursors, students with similar starting performance divided into trajectory classes with very different growth over time. Class 3 (*low-performers 1*) displayed an initial performance at about the same level as class 2, but growth was lower. Class 4 (*low-performers 2*) was highly interesting, as this group displayed the lowest starting performance but steep growth. Consequently, this class outperformed class 5 (*low-performers 3*, flat slope) and reached moderate scores at the end of the study.

## 3.3 Stability of trajectory group membership across competences

We examined to what degree students' classifications into trajectory classes were stable across the three competences. We were also interested in the consequences of being in one of the Basic Precursors or Advanced Precursors catch-up group (class 2 in both competences), i.e., whether students in these classes had increased odds of being in higher-performing or lower-performing classes in more advanced competences.

We first checked the contingency tables for general associations between the competences. Given that expected cell frequencies were below 5 for more than 20% of the cells, Fisher's exact test was used for this purpose. The test revealed significant associations (*p* < .001) for all three contingency tables (Basic Precursors × Advanced Precursors, Basic Precursors × Computation, Advanced Precursors × Computation). We then compared classifications across competences for specific classes and first analyzed the classification congruence of (relatively low-performing) class 3 Basic Precursors students. For these students, the odds of also being in a low-performing Advanced Precursors trajectory group (classes 3-5) were 14.43 times higher (95% CI [5.24, 34.41]) than for class 1 or class 2 Basic

Precursors students[4]. The classification of class 3 Basic Precursors students in Computation trajectory groups was also pronounced, with the odds of also belonging to a low-performing class (3-5) being 9.40 times higher (95% CI [3.73, 23.67]) than for class 1 or class 2 Basic Precursors students. In contrast, class 1 Basic Precursors students were very unlikely to be in a low-performing Advanced Precursors or Computation class (see Table 6 and Table 7 for observed and expected frequencies). Students in the Basic Precursors catch-up trajectory group (class 2) did not have meaningfully increased odds of belonging to low-performing or high-performing Advanced Precursors or Computation classes.

*Table 6.* Observed and expected Basic Precursors and Advanced Precursors class frequencies

| | | | Advanced Precursors | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 | Total |
| Basic Precursors | Class 1 | Count | 38 | 1 | 3 | 1 | 0 | 43 |
| | | Expected | 20.5 | 7.9 | 7.6 | 5.6 | 1.4 | 43.0 |
| | Class 2 | Count | 32 | 23 | 12 | 7 | 3 | 77 |
| | | Expected | 36.7 | 14.1 | 13.6 | 10.1 | 2.5 | 77.0 |
| | Class 3 | Count | 3 | 4 | 12 | 12 | 2 | 33 |
| | | Expected | 15.7 | 6.0 | 5.8 | 4.3 | 1.1 | 32.9 |
| | | Total | 73 | 28 | 27 | 20 | 5 | 153 |

*Table 7.* Observed and expected Basic Precursors and Computation class frequencies

| | | | Computation | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 | Total |
| Basic Precursors | Class 1 | Count | 28 | 11 | 2 | 1 | 1 | 43 |
| | | Expected | 14.1 | 12.1 | 9.8 | 3.4 | 3.7 | 43.1 |
| | Class 2 | Count | 20 | 27 | 20 | 6 | 4 | 77 |
| | | Expected | 25.2 | 21.6 | 17.6 | 6.0 | 6.5 | 76.9 |
| | Class 3 | Count | 2 | 5 | 13 | 5 | 8 | 33 |
| | | Expected | 10.8 | 9.3 | 7.5 | 2.6 | 2.8 | 33.0 |
| | | Total | 50 | 43 | 35 | 12 | 13 | 153 |

[4] Odds ratios were calculated from summated cell values. E.g., for the odds ratio of class 3 Basic Precursors students being in Advanced Precursors classes 3-5 compared to classes 1-2 Basic Precursor students being in Advanced Precursors classes 3-5, the ratio was (26/7) / (34/86) = 9.40. For the computation of confidence intervals, see Bland and Altman (2000).

The odds of students in the low-performing Advanced Precursors classes 3-5 of also being in low-performing Computation classes (classes 3-5) were similarly increased (odds ratio of 9.75, 95% CI [4.50, 21.13] compared to class 1 and class 2 Advanced Precursors students). Consequently, class 3-5 Advanced Precursor students were very unlikely to be in high-performing Computation classes, and only 2 of 52 students were in Computation class 1. The chances of belonging to the Computation catch-up group (class 2) were a little bit higher for these 52 students with low levels of Advanced Precursors. However, odds for this classification were still 4.46 times decreased (95% CI [1.91, 10.42]) compared to Advanced Precursors classes 1 and 2. As was the case with high-performing Basic Precursors students, class 1 Advanced Precursors students were very unlikely to be in low-performing Computation classes (see Table 8 for observed and expected frequencies). As was the case with class 2 Basic Precursors, class 2 Advanced Precursors students did not have meaningfully increased odds of belonging to low-performing or high-performing Computation classes.

*Table 8.* Observed and expected Advanced Precursors and Computation class frequencies

| | | | Computation | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 | Total |
| Advanced Precursors | Class 1 | Count | 41 | 21 | 9 | 1 | 1 | 73 |
| | | Expected | 23.9 | 20.5 | 16.7 | 5.7 | 6.2 | 73.0 |
| | Class 2 | Count | 7 | 10 | 8 | 3 | 0 | 28 |
| | | Expected | 9.2 | 7.9 | 6.4 | 2.2 | 2.4 | 28.1 |
| | Class 3 | Count | 2 | 6 | 11 | 3 | 5 | 27 |
| | | Expected | 8.8 | 7.6 | 6.2 | 2.1 | 2.3 | 27.0 |
| | Class 4 | Count | 0 | 4 | 7 | 4 | 5 | 20 |
| | | Expected | 6.5 | 5.6 | 4.6 | 1.6 | 1.7 | 20.0 |
| | Class 5 | Count | 0 | 2 | 0 | 1 | 2 | 5 |
| | | Expected | 1.6 | 1.4 | 1.1 | 0.4 | 0.4 | 4.9 |
| | | Total | 50 | 43 | 35 | 12 | 13 | 153 |

# 4

# Discussion

The current study extends the research on early mathematics learning by describing in detail first-grade growth trajectories across different competences. In essence, our findings support mainly cumulative growth patterns in all competences. I.e., performance at the beginning of the study predicted growth throughout the school year for the majority of students in that trajectory classes that display higher performance at time 1 also display higher performance at time 7. However, we also consistently found some compensatory growth patterns in all competences and overall scores. I.e., students with lower performance at time 1 subsequently followed diverse trajectory groups that were characterized by varying growth levels. This breakdown into different trajectory groups from a similar starting performance was particularly evident for the Computation competence, which assessed the main curricular goals of the school year.

## 4.1 Trajectory groups for overall mathematics achievement

Concerning our first two research question, LGCM revealed significant slopes and significant variance in students' performance level and slopes, indicating that meaningful learning took place from time 1 to time 7, during which students followed differing growth trajectories. LCGA revealed five trajectory groups for overall math scores which followed three distinct patterns. As was reported by previous longitudinal research (Aunola et al., 2004; Bodovski & Farkas, 2007; Geary et al., 2012; Jordan et al., 2007, 2006; Morgan et al., 2009; Morgan, Farkas, & Wu, 2011), the longitudinal pattern was mainly fan-spread. Essentially replicating findings by Jordan et al (2007, 2006), we found one additional class that started below average but then showed steeper growth than all other classes.

Somewhat different from Jordan et al.'s results, our data suggested two (essentially parallel) high-performing and two low-performing classes instead of just one each.

When interpreting these findings, one should keep in mind that the learning growth of the highest-performing group was limited by ceiling effects. The tasks in the progress monitoring tests mostly did not exceed a difficulty level that is expected at the end of first grade. High-performers might have shown higher growth if second-grade problems had been included in the test.

An important finding from overall score analyses is that two classes with more than half of the initially weak-performing students (27% of all students) showed too little growth to reach high scores by the end of the study. Nonetheless, all other classes (including a fairly large catch-up class with initially low performing students) reached high scores at the end of the study.

## 4.2 Trajectory groups for separate competences

Concerning our research question three, the overall fan-spread pattern with persistently high-performing and persistently low-performing classes was consistently found for the separate competences, as was a catch-up class with initially lower scores. This broad pattern showed specific characteristics for the different competences.

For Basic Precursors, more than one quarter of the children (class 1) showed ceiling effects from time 1, limiting learning growth over time, and 50% of all students belonged to a catch-up group, reaching very high scores towards the end of the study. For Advanced Precursors and Computation, classes with distinct compensatory patterns were smaller, but more pronounced. For Computation, three classes comprising almost 60% of the children started the study at very similar performance levels, but slopes ranged from completely flat to steep. These results are not surprising, considering that most Basic Precursors and many Advanced Precursors skills usually develop before the start of formal education (Krajewski & Schneider, 2009), but skills assessed in our Computation measures are curricular competences in first grade. Thus, most children have little prior knowledge in this competence and start at similar (low) performance levels.

## 4.3 Stability of trajectory group classifications

Our fourth research question concerned the homogeneity of group classifications across the competences. As expected, students in low-performing precursor classes were much less likely to be in high-performing Computation classes than were students from high-performing precursor classes. This pattern suggests that in order to reach the curricular computation goals, high precursor skills are essential. In addition to this general finding, students with low initial but steeply growing precursor performance did not have elevated outcome risks for higher-order competences. In other words, catch-up classes in Basic Precursors and Advanced Precursors did not have meaningfully increased odds of belonging to low-performing classes in higher-order competences.

## 4.4 Implications for research and practice

Our study has a number of implications for research and practice of first-grade mathematics over and above previous studies. First, our results indicate that a single assessment of students' skills at the beginning of formal schooling may not suffice to reliably identify students at risk of developing persisting math difficulties. Morgan et al. (2009) used cutoffs from two assessments at the beginning and end of kindergarten to determine the stability of math difficulties and found that only about half of the children identified as "at risk" at the first assessment displayed persistently low performance. The results of our study further suggest that the accuracy of risk estimations over and above cutoff values may be improved by using latent class categorizations.

Second, also concerning risk estimations, our study confirms previous findings that students with high overall performance at the start of the first school year have a high chance of also being high-performers at the end of the school year. Our study adds that it may be beneficial to examine different competences separately. Students with high initial Basic Precursors skills had a very high chance of mastering first-grade curricular goals even if they did not have considerable previous knowledge in Computation. For the majority of the students, their initial Computation performance did not seem to be the risk factor of choice, as three of the trajectory groups started at similar levels but then displayed very different learning growth. We therefore recommend to use precursor skills for risk estimations, as have numerous other authors (e.g., Chard et al., 2005; Krajewski & Schneider, 2009; Locuniak & Jordan, 2008). However, our study did not include long-term measures of math achievement, and further research is required to address the stability of these findings.

Third, concerning the question whether mathematical competence follows cumulative or compensatory patterns, our study provides evidence for both paths. In our sample, learning growth in all competences was mainly linear, suggesting high stability of the learning trajectories. We also observed that the gap between the highest-performing and the lowest-performing classes widened considerably within first grade. Nonetheless, our data suggests the presence of a catch-up class in every competence. These classes showed higher growth rates than the classes with the highest initial performance. However, growth rates of the highest-performing classes might have been limited due to ceiling effects.

Finally, we want to emphasize that we identified several important classes following the recommendation by Jung and Wickrama (2008) to consider interpretability of classes when conducting latent growth modeling. Selecting LCGA solutions with slightly poorer statistical fit resulted in visibility of the very poorly performing class 5 in Computation as well as the Advanced Precursors and overall score catch-up groups. These classes also showed distinct patterns in the analyses of group membership stability.

## 4.5 Limitations

At least two limitations have to be considered when interpreting the results of our study, which relate to the generalizability of the findings.

First, although our study sample covered schools in rural and urban areas and students had heterogeneous family backgrounds, generalizability of the results to other countries cannot be taken for granted. Nonetheless, the overall results pattern – students with high initial performance reach the curricular goals, students with initially low performance split into those with steep learning growth and others with little to no improvement over the school year – was also found in the studies by Jordan et al. (2007, 2006) in a culturally different setting.

Second, our study solely draws conclusions as to the performance trajectories within first grade. Studies which collected data over a longer time span mostly found more stable performance patterns (Aunola et al., 2004; Bodovski & Farkas, 2007; Geary et al., 2012; Jordan et al., 2007; Morgan et al., 2009). Some of these more stable results may be due to the analysis procedures used, but there is also evidence for generally stabilized performance in later elementary school grades (e.g., Geary et al., 2012; Kim, Petscher, Schatschneider, & Foorman, 2010).

## 4.6 Conclusions

Overall, the results of the present study suggest that first-graders have more diverse learning growth trajectories than suggested by multi-year longitudinal studies. The majority of students followed cumulative growth patterns, but some students showed very strong learning growth and high final scores at the end of the school year despite low initial performance. Analyses of class memberships across different competences revealed that students in precursor trajectory groups with low outcomes had a significantly elevated risk of not reaching curricular arithmetic goals at the end of the school year.

The results highlight the value of assessing diverse competences in first grade and stress the need for early intervention for students with precursor deficits. Closely monitoring students' progress is advisable to identify students who do not show sufficient learning growth (Espin et al., 2012; Fuchs & Fuchs, 1998).

# 5
# References

Aarnoutse, C., & van Leeuwe, J. (2000). Development of poor and better readers during the elementary school. *Educational Research and Evaluation*, *6*(3), 251–278. doi:10.1076/1380-3611(200009)6:3;1-A;FT251

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

Authors (in press). [Information masked for review purposes].

Bast, J., & Reitsma, P. (1997). Matthew effects in reading: A comparison of latent growth curve models and simplex models with structured means. *Multivariate Behavioral Research*, *32*(2), 135–167. doi:10.1207/s15327906mbr3202_3

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

Bland, J. M., & Altman, D. G. (2000). Statistics notes: the odds ratio. *BMJ: British Medical Journal*, *320*(7247), 1468.

Bodovski, K., & Farkas, G. (2007). Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction. *The Elementary School Journal*, *108*(2), 115–130. doi:10.1086/525550

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: a structural equation perspective*. New York, NY: Wiley.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, *20*(3/4/5/6), 487–506. doi:10.1080/02643290244000239

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(1), 128–141.

Espin, C. A., Mcmaster, K. L., Rose, S., & Wayman, M. M. (Eds.) (2012). *A measure of success: The influence of curriculum-based measurement on education*. Minneapolis: University of Minnessota Press.

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.

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research & Practice*, *13*(4), 204–219.

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities*, *37*(1), 4–15. doi:10.1177/00222194040370010201

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2012). Mathematical cognition deficits in children with learning disabilities and persistent low achievement: A five-year prospective study. *Journal of Educational Psychology*, *104*(1), 206–223. doi:10.1037/a0025398

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*, 549–76. doi:10.1146/annurev.psych.58.110405.085530

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. doi:10.1080/10705519909540118

Jordan, N. C., Kaplan, D., Locuniak, M. N., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice, 22*(1), 36–46. doi:10.1111/j.1540-5826.2007.00229.x

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

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass, 2*(1), 302–317. doi:10.1111/j.1751-9004.2007.00054.x

Kempe, C., Eriksson-Gustavsson, A.-L., & Samuelsson, S. (2011). Are there any matthew effects in literacy and cognitive development? *Scandinavian Journal of Educational Research, 55*(2), 181–196. doi:10.1080/00313831.2011.554699

Kim, Y.-S., Petscher, Y., Schatschneider, C., & Foorman, B. R. (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

Kolkman, M. E., Kroesbergen, E. H., & Leseman, P. P. M. M. (2013). Early numerical development and the role of non-symbolic and symbolic skills. *Learning and Instruction, 25*(0), 95–103. doi:10.1016/j.learninstruc.2012.12.001

Krajewski, K., Küspert, P., Schneider, W., Deimann, P., & Kastner-Koller, U. (2002). DEMAT 1+. Deutscher Mathematiktest für erste Klassen. [German mathematics test for first grades]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 34*(4), 236–238. doi:10.1026//0049-8637.34.4.238

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

Leppänen, U., Niemi, P., Aunola, K., & Nurmi, J.-E. (2004). Development of reading skills among preschool and primary school pupils. *Reading Research Quarterly*, *39*(1), 72–93. doi:10.1598/RRQ.39.1.5

Locuniak, M. N., & Jordan, N. C. (2008). Using kindergarten number sense to predict calculation fluency in second grade. *Journal of learning disabilities*, *41*(5), 451–9. doi:10.1177/0022219408321126

Mazzocco, M. M. M. (2005). Challenges in identifying target skills for math disability screening and intervention. *Journal of Learning Disabilities*, *38*(4), 318–323. doi:10.1177/00222194050380040701

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

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities*, *42*(4), 306–321. doi:10.1177/0022219408331037

Morgan, P. L., Farkas, G., & Wu, Q. (2011). Kindergarten children's growth trajectories in reading and mathematics: Who falls increasingly behind? *Journal of Learning Disabilities*, *44*(5), 472–488. doi:10.1177/0022219411414010

Muthén, B. O., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, *24*(6), 882–891.

Muthén, B. O. (2003). Statistical and substantive checking in growth mixture modeling: comment on Bauer and Curran (2003). *Psychological Methods*, *8*(3), 369–377. doi:10.1037/1082-989X.8.3.369

Muthén, L., & Muthén, B. O. (1998-2010). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a monte carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. doi:10.1080/10705510701575396

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J.-E., & Kirby, J. R. (2005). Development of individual differences in reading: Results from longitudinal studies in english and finnish. *Journal of Educational Psychology*. American Psychological Association. doi:10.1037/0022-0663.97.3.299

Phillips, L. M., Norris, S. P., Osmond, W. C., & Maynard, A. M. (2002). Relative reading achievement: A longitudinal study of 187 children from first through sixth grades. *Journal of Educational Psychology, 94*(1), 3–13. doi:10.1037/0022-0663.94.1.3

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Schlomer, G. L., Bauman, S., & Card, N. a. (2010). Best practices for missing data management in counseling psychology. *Journal of counseling psychology*, *57*(1), 1–10. doi:10.1037/a0018082

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

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407. doi:10.2307/747612

Williamson, G. L., Appelbaum, M., & Epanchin, A. (1991). Longitudinal analyses of academic achievement. *Journal of Educational Measurement*, *28*(1), 61–76. doi:10.1111/j.1745-3984.1991.tb00344.x

# PART IV

## Manuscript 3
## Web-based mathematics progress monitoring in second grade

# Web-based mathematics progress monitoring in second grade

Martin Salaschek[a] & Elmar Souvignier[a]

[a] University of Münster, Germany

Corresponding author:
Martin Salaschek, Institute for Psychology in Education, University of Münster,
Fliednerstrasse 21, 48149 Muenster, Germany
Email: salaschek@uni-muenster.de

Word count

| | |
|---|---|
| Abstract: | 131 words |
| Main text: | 4700 words |
| References: | 1108 words |
| Tables & Figures: | 724 words |

## Abstract

We examined a web-based mathematics progress monitoring tool for second-graders. The tool monitors the learning progress of two competences, number sense and computation. 414 students from 19 classrooms in Germany were checked every three weeks from fall to spring. Correlational analyses indicate that alternate form reliability was adequate for the chosen interval (.81 < *r* < .87). Results from latent growth curve modeling (LGCM) identified significant linear increases in students' scores, and significant variance in computation slopes was observed. Criterion validity coefficients, comparing the measures with student performance on standardized school achievement tests at the beginning and the end of the school year (DEMAT1+, DEMAT2+), were satisfactory (.59 < *r* < .76). Students conducted the tests independently, and assessment times were short. Implications for further research and classroom practice are discussed.

*Keywords:* progress monitoring, mathematics, assessment, number sense

# 1 Introduction

Significant differences between children's mathematics skills can be observed at and even before the beginning of formal schooling. These differences are strong predictors of later achievement (Bodovski & Farkas, 2007; Jordan, Kaplan, Oláh, & Locuniak, 2006; Missall, Mercer, Martínez, & Casebeer, 2012). Facing these individual differences, it has been shown that learners at all performance levels profit when teachers individualize their instruction based on students' strengths and weaknesses (Connor, Morrison, & Petrella, 2004; Stecker & Fuchs, 2000). Furthermore, at-risk students should be identified early so that suitable interventions can be implemented, and a need for modified or additional instruction can be derived from low performance in *number sense* competences (Berch, 2005). There is also evidence that students with similar initial number sense competences divide into trajectory groups with either flat or steep learning growth (Jordan, Kaplan, Locuniak, & Ramineni, 2007; Jordan et al., 2006). Thus, in addition to one-time screenings, teachers need to document students' progress to identify those who show a lack of learning gains.

Progress monitoring tools can assist teachers in determining students' specific skills and their learning progress, and they can assist in drawing conclusions for instructional adjustments (Allinder & Beckbest, 1995; Stecker, Fuchs, & Fuchs, 2005; Stecker & Fuchs, 2000). These tools should therefore reliably assess students' performance and progress. Furthermore, they should provide teachers with instructionally relevant information about various curricular competences for all students and should be as effortless as possible in their implementation such that general classroom work is not hindered (Clarke, Baker, Smolkowski, & Chard, 2008; Förster & Souvignier, 2011). The aim of the present study was to examine a newly-developed progress monitoring tool consistent with these goals for second grade mathematics.

## 1.1 Progress monitoring in elementary school mathematics

Progress monitoring has a long history especially in the U.S., where curriculum-based measurement (CBM) as a form of progress monitoring was introduced in special education more than 30 years ago (see Deno, 2003, for an overview). In CBM, usually very short measures (e.g., solving as many addition problems as possible in one minute) are conducted weekly to assess students' performance development over time. In addition to meeting typical test criteria—such as reliability and validity—CBM tests need to assess skills that progress over time so that score increases can be expected between the tests. Depending on the individual rate of growth, instruction can then be adjusted if deemed necessary. As a prerequisite for this sensitivity to learning growth, all test forms need to be parallel so that increasing scores indeed can be interpreted as learning growth.

CBM tests usually produce very high alternate-form or test-retest reliabilities of .80 and higher. Where reported, criterion validity scores are also often satisfactory, frequently exceeding .50. Many tests are also sensitive to students' learning such that average scores slightly increase from week to week (see Foegen, Jiban, and Deno, 2007, for an overview). While CBM is widely used in special education or to determine eligibility for special education, some restrictions of traditional CBM tests have to be overcome for a use in general education, directed towards individualized instruction for all students. The foremost obstacle seems to be time constraints of general education teachers (Deno, 2003). Given that CBM tests usually need to be individually scored by the teachers, conducting weekly tests of whole classrooms seems laborious.

Another obstacle concerns the use of the results for individualized instruction. In elementary school, most math CBM tests for grade 2 and higher consist of only one skill domain, namely basic arithmetic problems (Foegen et al., 2007). While these measures produce very reliable scores, they do not provide teachers with information on specific strengths and weaknesses of their students which are needed for individualized instruction. Christ and colleagues (Christ, Scullin, Tolbize, & Jiban, 2008, p. 204) argued that single-skill computation assessments "should not be interpreted to represent mathematics achievement generally". The usefulness of single-skill tests for improving instruction thus has rarely been documented (Fuchs, 2004).

In contrast, tests that include multiple measures to assess a variety of relevant abilities seem more adequate for generalized and longer-term performance (Hintze, Christ, &

Keller, 2002). However, including multiple measures impedes high reliability levels of a test, especially with short assessment times. Christ, Johnson-Gros, and Hintze (2005) consequently reasoned that 10-15 minutes test time should be appropriate for tests assessing several curricular measures.

We therefore strived to design a progress monitoring tool that includes a number of different aspects of mathematical competence, sufficient to draw conclusions for classroom work. To ensure the tool's feasibility, implementation should be effortless, and reliability must be high enough to allow a dependable assessment of student progress.

## 1.2 Purpose of the study

The purpose of this study was to analyze a newly-developed progress monitoring tool for general-education math skills that provides teachers with information about students' performance status along with their learning progress throughout the school year. The test included measures of number sense competence for a rating of basic mathematical understanding as well as curriculum-based measures of computation competence for a detailed assessment of second-grade skills.

Research questions are as follows: (1) How reliable are the scores obtained from our progress monitoring tool? (2) Is the tool sensitive to student learning? I.e., can increases in scores over time be observed? (3) How do the test scores relate to standardized school achievement test results and to teacher ratings of students' mathematical performance at the beginning and end of the school year? (4) Is the feasibility high enough so that teachers' and students' acceptance of the tool is achieved?

# 2
# Method

## 2.1 Participants and setting

414 students (212 male) from 19 second-grade general education classrooms in Germany participated in the study. 17% of the students spoke a foreign language at home. The average age of students at pretest was 7.60 years (*SD* = 0.57). Participating schools were located in urban and rural areas of Germany.

The study was conducted from October 2011 to June 2012. In October 2011, the paper pencil test DEMAT1+ was administered. DEMAT1+ was immediately followed by progress monitoring tests every three weeks, running from October to May. At the end of the study, a second paper pencil test was administered, the DEMAT2+. Teachers were asked to rate their students' mathematical competence before both paper pencil tests.

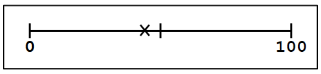
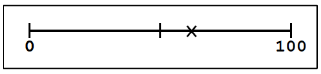
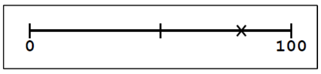
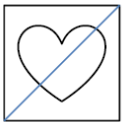
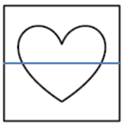
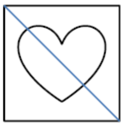
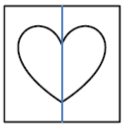
Research question 4 was explored in a congruous study with 13 second-grade teachers in the 2010/2011 school year. After the last progress monitoring test, teachers were surveyed about their use of the progress monitoring tool. Questions concerned progress monitoring implementation and teachers' use of the results for classroom purposes. In that school year, progress monitoring tests were conducted every two weeks (instead of every three weeks), and single test items were revised after that school year. All other progress monitoring procedures were the same for both school years.

## 2.2 Measures and procedure

### 2.2.1 Progress monitoring measures

The progress monitoring test concept included two different competences: number sense and computation. Both competences contained several different types of tasks (referred to as "measures" in the following, to discern them from single questions and competence scores), which included several problems each. Table 1 provides an overview of the test concept. The measures included in the competences were selected with regard to their theoretical importance for mathematical understanding and their relevance in the curriculum.

*Table 1.* Overview of Progress Monitoring Measures

| Measure | Number of items | Example problem and distractors |
|---|---|---|
| *number sense* | *24* | |
| number recognition | 8 | *Audio:* "72" … 82 \| 27 \| 72 \| 26 |
| size comparison | 6 | 15€ 38ct … < \| = \| > … 39€ 10ct |
| number line | 6 | *Audio:* "81"<br> <br> <br> |
| axis of symmetry | 4 |     |
| *computation* | *28* | |
| addition | 4 | 26 + 22 = … 46 \| **48** \| 58 \| 56 |
| subtraction | 4 | 72 - 23 = … 57 \| 47 \| **49** \| 59 |
| multiplication | 4 | 4 x 3 = … 10 \| 14 \| **12** \| 11 |
| double | 6 | 13 … 23 \| **26** \| 36 \| 24 |
| divide in half | 6 | 30 … **15** \| 10 \| 25 \| 20 |
| add up to 100 | 4 | 21 … 79 \| 87 \| 77 \| **89** |

*Note.* All measures contained items of varying degree of difficulty.

The first competence, number sense, served as an indicator for fundamental mathematical understanding and was based on the *triple-code model* (Dehaene, 1992, 2001, 2011; Dehaene & Cohen, 1995) and Krajewski's *model of early mathematical development* (Krajewski, 2008;

Krajewski & Schneider, 2009). Four different measures were included in the competence, three of which closely relate to the two models.

In *number recognition*, students were required to identify numbers in the range of 1-1000 after hearing the number via headphones. This measure tested the link between verbal number representations and Arabic numbers as described in the triple-code model.

In *size comparison*, students had to choose the correct equality or inequality operator (>, <, =) based on two amounts of money displayed on the screen. This procedure tested the precise quantity-number link as described in Krajewski's model.

*Number line* items consisted of problems where numbers in the range of 1-100 were given to students via headphones. Students were required to identify the number line on the screen marked with this number against several distractor number lines. The number lines were marked with 0, 100 and a blank mark in the middle of the line. This procedure assessed the development of a mental number line as proposed in the triple-code model's analogue magnitude representation system.

*Symmetry* items required students to identify the correct axis of symmetry for geometric shapes. Symmetry items were included in the test because of their importance in most elementary school curricula worldwide (e.g., National Council of Teachers of Mathematics, 2012). Clements (2003) cited evidence that many geometric skills, including a sense for symmetry, develop even before school entrance. Thus, the measure was categorized as a number sense measure.

Generally, skills necessary for mastering these measures start to develop before or during first grade. However, children reach higher competence levels for small quantities first, and the development for larger quantities can lag several years behind (Krajewski & Schneider, 2009). Thus, all number sense measures included problems with quantities ranging from small to large.

The second competence, computation, aimed to assess second-grade curricular goals. The competence comprised six different measures with arithmetic problems in the range of 1-100. Arithmetic proficiency in this range forms the core curricular goal in second grade (compared to proficiency in the range of 1-20 in first grade).

In *add up to 100*, a number was presented for which a missing summand had to be found so that both numbers summed to 100. This measure tested students' ability to compose and decompose large numbers, an extension of the highest level in Krajewski's model of early mathematical development.

*Doubling* and *dividing in half* required students to double or divide the presented number in half and pick the correct solution from several distractors. These measures tested students' basic conceptualization of multiplication and division.

Finally, *addition*, *subtraction*, and *multiplication* problems required students to pick the correct solution for a conventional arithmetic problem from several distractors.

Again, problems with quantities ranging from small to large were included in the measures to target diverse skill levels.

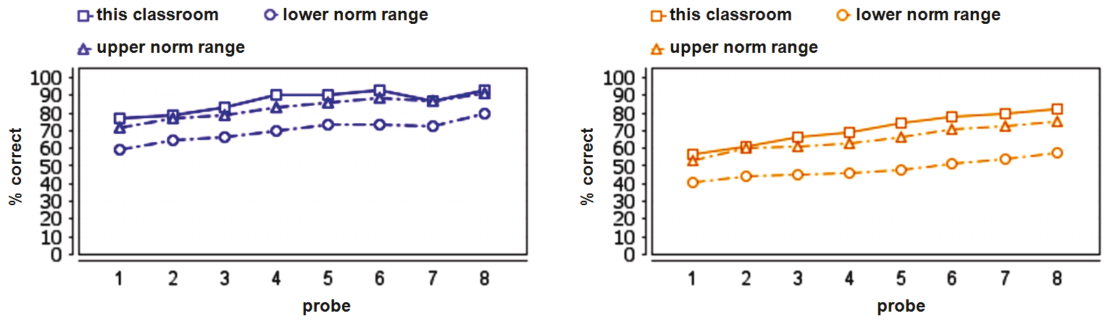
### 2.2.2 Progress monitoring procedure

Four parallel versions A-D of the test were created, and each was completed twice (sequence A-D, A-D) so that students completed eight tests in total. To achieve parallelism, all test items were carefully designed following item cloning strategies (e.g., keeping the distance of minuend and subtrahend as well as the overall magnitude of a result constant; cf. Clause, Mullins, Nee, Pulakos, & Schmitt, 1998).

Every three weeks students in all classrooms completed a test. A time frame of three weeks was chosen to attain balance between the density of diagnostic information and practicality.

To facilitate the administration and processing of the test and its results for the teachers, all progress monitoring activity was computerized. Progress monitoring tests were presented to the students online in multiple-choice format (clickable pictures). Audios, which included a general introduction to the test, explanations before each measure, and parts of the task in some of the measures were provided via headphones. Students worked on the tests independently and the tests were self-paced. Where a classroom had computer rooms available for testing, all students completed a test at the same time. Elsewhere, students completed the tests during self-study periods in the classroom.

Test scores were automatically calculated as percentage of correct answers in each competence. Results for both competences could be examined separately and online by the teachers directly after the test as graphs and tables. Results could be viewed at the student-level and class-level. Reference values (i.e., mean values of all participating classes, including +/ – 1 SD) could be added to the results view at class-level. Figure 1 shows a sample screenshot of a teacher's view with study-wide comparisons enabled.



| | | probe 1 | probe 2 | probe 3 | probe 4 | probe 5 | probe 6 | probe 7 | probe 8 |
|---|---|---|---|---|---|---|---|---|---|
| score (norm) | number sense | 76.9<br>(67.9) | 78.3<br>(71.8) | 83.1<br>(74.0) | 90.0<br>(78.2) | 90.4<br>(81.0) | 92.7<br>(82.6) | 83.5<br>(81.4) | 87.1<br>(84.8) |
| | curriculum-based | 55.6<br>(46.5) | 58.9<br>(51.6) | 65.9<br>(52.5) | 69.3<br>(54.3) | 74.5<br>(58.1) | 78.0<br>(63.1) | 79.3<br>(64.3) | 82.0<br>(65.3) |

*Figure 1.* Teacher's view of results (class level, study-wide comparisons enabled).

### 2.2.3 Criterion measures

The two school achievement tests preceding and following the progress monitoring tests were standardized paper pencil tests. DEMAT1+ (Krajewski, Küspert, Schneider, Deimann, & Kastner-Koller, 2002) was used as a measure of concurrent validity. The assessment is suitable for late first grade and early second grade. It mainly contains tasks from first-grade curricula (e.g., addition and subtraction in the range from 1-20) but also includes some items testing for number sense (magnitudes and numbers, number ranges). In its reference sample for second grade, DEMAT1+ showed high internal consistency (α = .88) and correlation with teacher ratings (*r* = .66).

At the end of the school year, online tests were followed by DEMAT2+ (Krajewski, Liehm, & Schneider, 2004), suitable for late second grade and early third grade. DEMAT2+ was used as a measure of predictive validity. The assessment mainly contains tasks from second-grade curricula (e.g. arithmetic problems in the range of 1-100. In its reference sample for second grade, DEMAT2+ showed high internal consistency (α = .93) and correlation with end-of-year grades (*r* = .66) along with predictive validity of third-grade (*r* = .65) performance.

Before each paper pencil test, teachers were asked to rate their students' overall mathematical abilities on a 7-point Likert scale ("The student's overall mathematics skill level is [*very weak* to *very strong*]").

# 3 Results

After the presentation of descriptive statistics, results for the research questions are explored in four steps. First, the tests' reliability is analyzed via correlations of test results across the different probes. Second, the tool's capacity to model learning growth is examined via latent growth curve modeling (LGCM; Bollen & Curran, 2006). Third, criterion validity is analyzed by reviewing how the progress monitoring results relate to paper pencil test results and teacher ratings. Finally, results from the teacher survey on feasibility and usage are reported.

LGCM as well as correlation analyses were conducted within a multi-level framework using Mplus 7.11 (Muthén & Muthén, 1998-2010), taking into account school class membership of students. Following this procedure, all correlation coefficients were the same as in single-level analyses, but standard errors take into account the multi-level structure (Cohen, Cohen, West, & Aiken, 2003).

Some test data were missing because of students being sick during a test or leaving the school for other reasons throughout the school year. Missing data ranged from 1-6% for any of the eight progress monitoring tests. For the pretest and posttest, 2% and 3% of the data were missing, respectively. These missing data were handled by means of full information maximum likelihood (FIML; Enders, 2001). FIML requires that data are missing at random (MAR). MAR can be assumed if variables are included in the data set which closely relate to the variables containing missing data (Collins, Schafer, & Kam, 2001). Given the number of strongly correlated variables in our study, we assumed that data was MAR. In addition, 14% and 2% of the teacher ratings of their students' skills had not reached us in timely manner and were thus also declared missing. For analyses involving teacher ratings, only students with complete teacher ratings present were included in the analysis (*n* = 347) because ratings were missing for complete classrooms, and not enough relating data was available to assume MAR.

Scores and standard deviations of the progress monitoring competences and overall scores[1] are depicted in Table 2. Overall, 56.4% of the problems were answered correctly at test 1 and 74.3% at test 8. A larger percentage of number sense answers than computation answers was correct. With few exceptions, competence and overall scores increased from test to test.

*Table 2.* Descriptive Statistics of progress monitoring overall score and competences

| | overall score | | number sense | | computation | |
|---|---|---|---|---|---|---|
| items | 52 | | 24 | | 28 | |
| time | M | *SD* | M | *SD* | M | *SD* |
| 1 | 29.33 | 8.49 | 16.30 | 4.44 | 13.03 | 5.02 |
| 2 | 31.68 | 8.89 | 17.23 | 4.61 | 14.45 | 5.29 |
| 3 | 32.47 | 8.84 | 17.76 | 4.24 | 14.71 | 5.60 |
| 4 | 33.97 | 9.02 | 18.76 | 4.28 | 15.21 | 5.76 |
| 5 | 35.71 | 8.83 | 19.44 | 4.10 | 16.28 | 5.65 |
| 6 | 37.53 | 8.90 | 19.85 | 4.11 | 17.68 | 5.69 |
| 7 | 37.54 | 8.72 | 19.53 | 3.75 | 18.01 | 5.87 |
| 8 | 38.62 | 9.55 | 20.33 | 4.17 | 18.29 | 6.32 |

## 3.1 Reliability

We computed Cronbach's α for total scores and both competences as a measure of internal consistency. Consistencies of total scores were high (.91 ≤ α ≤ .96 for the eight tests, *M* = .93). Number sense (.79 ≤ α ≤ .90, *M* = .85) and computation (.81 ≤ α ≤ .91, *M* = .86) also demonstrated good consistency.

Delayed alternate-form reliability was calculated for each adjacent test (t1×t2, t2×t3 … t7×t8), resulting in seven comparisons per competence. Correlation indices ranged from .71 to .79 (*M* = .76) for number sense competence and .75 to .81 (*M* = .78) for curriculum-based competence. For overall test scores, adjacent-test correlations ranged from .81 to .87 (*M* = .84). With increasing time between the tests (e.g., test 1 × test 4), the statistical connection decreased slightly (Table 3).

[1] To ascertain dimensionality, we conducted latent confirmatory factor analyses for all eight time points in Mplus. Results indicate that two-factor models with number sense and computation as factors fit the data for all time points (mean fit indices: CFI: 0.956; TLI: 0.942; RMSEA: 0.056; SRMR: 0.037).

*Table 3.* Correlations of progress monitoring overall scores

| | 2. (*SE*) | 3. (*SE*) | 4. (*SE*) | 5. (*SE*) | 6. (*SE*) | 7. (*SE*) | 8. (*SE*) |
|---|---|---|---|---|---|---|---|
| 1. overall score 1 | .81 (.019) | .79 (.023) | .75 (.027) | **.75 (.023)** | .71 (.031) | .70 (.032) | .66 (.033) |
| 2. overall score 2 | | .83 (.018) | .81 (.019) | .78 (.023) | **.76 (.025)** | .77 (.019) | .71 (.023) |
| 3. overall score 3 | | | .84 (.018) | .82 (.019) | .78 (.021) | **.79 (.021)** | .75 (.023) |
| 4. overall score 4 | | | | .83 (.028) | .78 (.032) | .78 (.027) | **.78 (.026)** |
| 5. overall score 5 | | | | | .81 (.037) | .81 (.022) | .81 (.021) |
| 6. overall score 6 | | | | | | .86 (.015) | .83 (.022) |
| 7. overall score 7 | | | | | | | .87 (.021) |
| 8. overall score 8 | | | | | | | |

*Note.* Correlations and standard errors of same tests are printed in bold. $p < .001$ for all correlations.

## 3.2 Sensitivity to learning

Overall sensitivity to learning was determined by LGCM. Separate models were computed for overall scores and each competence. Fit indices for models with and without quadratic slope estimates are reported in Table 4. Fit can be regarded as good to satisfying for all models (Hu & Bentler, 1999; Schermelleh-Engel, Moosbrugger, & Müller, 2003). Given that model fit was slightly higher for models taking quadratic slope into account, these models were then further evaluated. The models explained a large proportion of variance in each time point. For overall scores, R$^2$ varied from .80 to .88 ($M$ = .83). These values were slightly lower for number sense (.69 ≤ R$^2$ ≤ .81, $M$ = .75) and computation (.74 ≤ R$^2$ ≤ .83, $M$ = .78). Estimates of means and variances are reported in Table 5. Large variance across the estimated intercepts suggests significant differences in students' initial scores. Additionally, considerable linear growth over time is indicated by linear slope coefficients being significantly higher than zero. Quadratic slope coefficients below zero suggest slightly decelerating growth, but these coefficients were not significantly different from zero. The variance in slopes was significant for the linear slope in computation only. This result indicates that children did not differ significantly in learning growth for number sense and overall scores, but did so in computation.

*Table 4.* Fit indices for overall scores and competences

| | parameters | $\chi^2$ (df, n) | CFI | TLI | RMSEA | SRMR |
|---|---|---|---|---|---|---|
| overall scores | i, s | 64.99 (31, 414)*** | 0.976 | 0.978 | 0.071 | 0.054 |
| | i, s, q | 64.99 (27, 414)*** | 0.986 | 0.985 | 0.058 | 0.049 |
| number sense | i, s | 111.78 (31, 414)*** | 0.963 | 0.967 | 0.079 | 0.063 |
| | i, s, q | 65.84 (27, 414)*** | 0.982 | 0.982 | 0.059 | 0.055 |
| computation | i, s | 85.55 (31, 414)*** | 0.978 | 0.980 | 0.065 | 0.057 |
| | i, s, q | 74.22 (27, 414)*** | 0.981 | 0.980 | 0.065 | 0.053 |

*Note.* CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual. *** $p$ < .001

*Table 5.* LGCM estimates for intercepts and slopes of overall and competence scores

| | overall scores | number sense | computation |
|---|---|---|---|
| intercept (SE) | 29.43 (0.565)*** | 16.27 (0.250)*** | 13.15 (0.355)*** |
| var. of intercept (SE) | 61.04 (4.799)*** | 14.61 (1.091)*** | 20.44 (2.146)*** |
| slope (SE) | 1.86 (0.155)*** | 1.00 (0.099)*** | 0.89 (0.106)*** |
| var. of slope (SE) | 0.80 (0.529) | 0.32 (0.195) | 0.79 (0.359)* |
| quadratic slope (SE) | -0.08 (0.019) | -0.06 (0.012) | -0.02 (0.013) |
| var. of quadratic slope (SE) | 0.01 (0.009) | 0.00 (0.004) | 0.01 (0.005)+ |

*Note.* var. = variance. + $p$ < .10. * $p$ < .05. *** $p$ < .001

## 3.3 Criterion validity

### 3.3.1 Concurrent validity

Correlation coefficients between progress monitoring test scores (each competence and total scores) and the DEMAT1+ pretest were calculated. Results are displayed in Table 6. The correlations of progress monitoring overall scores with the first paper pencil test were moderate to strong and very stable across the tests (mean correlation of $r$ = .62). The coefficients for number sense were slightly higher than the coefficients for computation in this analysis.

### 3.3.2 Predictive validity

Correlations of progress monitoring overall scores with DEMAT2+ scores were strong (mean correlation of $r = .75$). It is noteworthy that correlation of both number sense and computation with DEMAT2+ (means being $r = .67$ and $r = .68$, respectively) were on the same level as the correlation of DEMAT1+ with DEMAT2+ ($r = .67$). Furthermore, correlations of progress monitoring overall scores with DEMAT2+ ($.72 \leq r \leq .77$) exceeded the correlations of DEMAT1+ with DEMAT2+ in all tests.

*Table 6.* Correlations of progress monitoring overall scores with paper pencil tests

| | DEMAT 1+ | | | DEMAT 2+ | | |
|---|---|---|---|---|---|---|
| | total scores | number sense | computation | total scores | number sense | computation |
| time | *r* (SE) | *r* (SE) | *r* (SE) | *r* (SE) | *r* (SE) | *r* (SE) |
| 1 | .59 (.028) | .55 (.035) | .51 (.029) | .73 (.027) | .66 (.035) | .65 (.027) |
| 2 | .60 (.038) | .60 (.043) | .49 (.034) | .72 (.028) | .65 (.030) | .64 (.028) |
| 3 | .60 (.037) | .54 (.055) | .54 (.030) | .76 (.027) | .66 (.034) | .70 (.027) |
| 4 | .63 (.041) | .60 (.048) | .55 (.036) | .74 (.029) | .67 (.039) | .66 (.030) |
| 5 | .63 (.035) | .61 (.036) | .54 (.037) | .75 (.025) | .68 (.024) | .68 (.030) |
| 6 | .62 (.041) | .59 (.043) | .54 (.042) | .75 (.026) | .66 (.041) | .70 (.023) |
| 7 | .63 (.041) | .58 (.048) | .56 (.039) | .77 (.027) | .69 (.034) | .69 (.032) |
| 8 | .63 (.037) | .62 (.036) | .54 (.039) | .77 (.021) | .67 (.033) | .72 (.021) |

*Note.* Correlation of DEMAT 1+ with DEMAT 2+ was r = .67. $p < .001$ for all correlations.

### 3.3.3 Teacher ratings

Overall progress monitoring scores were correlated with teachers' ratings of their students' mathematical abilities. Results show moderate correlations of the ratings provided before DEMAT1+ with the eight progress monitoring test scores ($.57 < r < .61$). Correlations of the second rating (before DEMAT2+) with the progress monitoring scores were considerably higher ($.66 < r < .70$).

## 3.4 Progress monitoring feasibility and usage

The 13 teachers participating in the survey were mainly positive about the progress monitoring implementation. Twelve of the 13 teachers would recommend the program to fellow teachers and t stated that the children were able to complete the tests independently (92% agreement each).

Teachers further declared that they used the CBM results diversely for classroom purposes. Apart from obtaining general information about the students and the class performance (85%, 79% agreement, respectively), teachers found the information especially useful when they were previously unsure of a student's performance (69% also used the system for this purpose). Teachers claimed to have at least sometimes given adjusted exercises based on CBM test results (77%, 69% agreement, respectively). A majority of respondents also found the information useful for designing supplementary education (54% agreement) or communicating about performance with students, parents and fellow teachers (85% agreement). The main concern of several teachers participating in the 2010/2011 school year project was the two-week time frame per test. They wished for longer testing intervals to allow more time for analyzing and working with the results.

As a direct measure of feasibility, the time needed to complete a test was recorded for each student. The median time needed to complete the first test, including all instructions, was 15.62 min. Subsequent median test times were considerably lower and declined constantly throughout the remaining tests (from 11.73 min for test 2 to 8.07 min in test 8; $M$ = 10.03 min). The difference between the first test and all other tests was partly due to initial starting instructions to the test (approx. 1 minute).

# 4

# Discussion

The purpose of the present study was to analyze a newly-developed assessment tool which overcomes typical barriers for the use of progress monitoring in general education. The tool comprised two competences, each based on multiple measures. To have utility, the assessment should reliably inform educators about diverse aspects of students' (static) performance and their development over time.

Several tests of reliability and validity explored psychometric properties of the static scores. Internal consistencies for overall scores and single competences were adequate. Moreover, we found adjacent test correlations to be strong for total scores, and correlations of the two single competences were only slightly lower. Therefore, results suggest that students' performance was reliably assessed.

As measures for criterion validity, correlations of the progress monitoring tests with DEMAT1+ scores were adequate, but considerably lower than correlations with DEMAT2+. Given that DEMAT1+ mainly assesses first-grade skills and DEMAT2+ assesses second-grade skills, the pattern reflects the orientation of our test concept towards second-grade competences. This assumption is also consistent with the finding that DEMAT1+ correlations were higher for number sense than for (second-grade) computation. As a general measure of predictive validity, we found that the association of each single progress monitoring test with DEMAT2+ was strong. The DEMAT2+ is representative of all German second-grade math curricula. Thus, even a single probe of the progress monitoring test early in the school year is a good indicator of the end-of-year performance.

In addition to psychometric properties of the static scores, significant positive linear growth for both competences in LGCM analyses indicate that the tests are sensitive to student learning. Observing quadratic (decelerating) slopes added to the model fit, but quadratic terms of the slopes were not significantly different from zero, ascertaining that the pattern was indeed mostly linear. The quadratic proportion might have been caused by "saturation effects": As students acquired new skills, the initial learning progress was fast, but it slowed down when students had established basic principles of the skill.

Regarding feasibility, assessment times were short and teachers regarded the progress monitoring tool as positive. Most participants evaluated the software as easy-to-use for the students and the results as useful for the assessment of students' competence and for making instructional changes. Although the teacher survey was conducted in a previous school year, the results pattern seems to be stable for the assessment system: Very similar results were obtained for the same assessment procedure with teachers in different grades and also in reading (Authors, in press; Authors, 2011).

## 4.1 Limitations

Our test concept—assessing diverse skills in complete general-education classrooms—differs considerably from typical progress monitoring assessments like CBM. Therefore, results cannot be compared directly to these application scenarios. While reliability in our study matches the coefficients typically found in CBM, test intervals of three weeks (instead of usually one week in CBM; Fuchs, 2004) might affect the reliability of slope estimations. However, Jenkins, Graff, and Miglioretti (2009) found that a three-week interval did not result in less accurate group-level growth estimates than weekly tests if the baseline score comprised several measures, and slope variances in our study were comparable to what is typically found in CBM studies (Foegen et al., 2007).

Another limitation of this study concerns the confirmation of parallelism of the test forms as an additional requirement for progress monitoring tests. No direct measure for parallel-forms reliability could be obtained in our study, given that different test forms were not administered at the same time. Yet, strong and narrow-ranging adjacent-test correlations suggest rank-order parallelism. In addition, descriptive statistics and LGCM analyses demonstrated mostly linear growth of scores which, given that the four test forms were each conducted twice, further hints parallelism. If the test forms had different difficulty levels, then dips or peaks in test scores should have occurred.

Finally, although teachers stated that they used the results for adjustments in classroom work or for designing supplementary interventions, we cannot describe the effects on teachers' instructional decision-making in detail because we relied on self-report data and did not perform classroom observations. Nonetheless, the increase in correlations between teacher ratings of students' competence and the test scores from the beginning to the end of the study can be regarded as an indicator of teachers recognizing and dealing with the data on learning growth. This result of increasing judgment accuracy of teachers is in line with meta-analytic findings from Südkamp, Kaiser, and Möller (2012), investigating uninformed and informed accuracies of teacher judgments of their students' academic performance. Stecker, Fuchs, and Fuchs (2005) reviewed research on the effect of progress monitoring usage on student achievement and drew the conclusion that significant additional growth in student learning could be seen when teachers used skills analyses for instructional modifications. Yet, this additional learning growth could only be observed when teachers earnestly utilized the results of a skills analysis. Mere administration of computer-based CBMs and sighting of the results by teachers failed to enhance learning outcomes. This result suggests that knowledge of students' performance level alone is not sufficient to improve teaching, and more information about the *quality* of students' knowledge is needed for improvement to occur.

Applying this finding to the progress monitoring tool at hand, its implementation can be expected to enhance student learning if teachers use the results to analyze students' skills and adjust instruction accordingly. We attempted to decrease barriers for such use by providing information with different granularity, as in the following: (1) Total scores allow overall classification of a student's performance; (2) number sense and computation indicate whether students have acquired basic skills and whether working on subsequent curricular goals is reasonable; (3) Single measure scores can be used to analyze strengths and weaknesses of a student and plan individualized instruction.

## 4.2 Implications for research and practice

Although our study provides evidence that the newly-developed progress monitoring tool reliably assesses students' math competences and that it is sensitive to student learning, further research is required on how data-driven classroom decisions can improve student learning, e.g., based on observed slopes (Ardoin, Christ, Morena, Cormier, & Klingbeil, 2013) or on analyses of single skills (Stecker et al., 2005; see also Shapiro, 2013). The presence of substantial slope standard errors compared to the average slopes represent substantial differences in students' learning growth, which advocates the use of progress monitoring to identify students who do not progress at a satisfactory rate. Identifying students who follow favorable or unfavorable learning paths may be aided by longitudinal studies using growth modeling to recognize students with distinct trajectory groups. Jordan and colleagues (2006) used growth mixture modeling for this purpose and found three trajectory groups: While a high-performers group had the highest initial and end-level performance in kindergarten, two groups performed similarly low at the beginning of the year but differed substantially in their growth. One of these groups displayed flat growth throughout the year, the other group had moderate to steep growth. For the purpose of providing students with individually adequate support it thus seems reasonable to broaden our knowledge on learning trajectories.

In conclusion, teachers can document student performance in general education throughout the school year at the classroom and individual level with the progress monitoring tool at hand. Test times are short, and there is no need for face-to-face assessment or manual scoring because test conduction and the results view are fully automated. Further research is required on how this progress monitoring data can be systematically used to improve student learning.

# 5

# References

Allinder, R. M., & Beckbest, M. A. (1995). Differential effects of two approaches to supporting teachers' use of curriculum-based measurement. *School Psychology Review*, *24*(2), 287–298.

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. doi:10.1016/j.jsp.2012.09.004

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

Authors (in press). [Information masked for review].

Authors (2011). [Information masked for review].

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

Bodovski, K., & Farkas, G. (2007). Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction. *The Elementary School Journal*, *108*(2), 115–130. doi:10.1086/525550

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: a structural equation perspective*. New York, NY: Wiley.

Christ, T. J., Johnson-Gros, K. N., & Hintze, J. M. (2005). An examination of alternate assessment durations when assessing multiple-skill computational fluency: The generalizability and dependability of curriculum-based outcomes within the context of educational decisions. *Psychology in the Schools*, *42*(6), 615–622.

Christ, T. J., Scullin, S., Tolbize, A., & Jiban, C. L. (2008). Implications of recent research: Curriculum-based measurement of math computation. *Assessment for Effective Intervention*, *33*(4), 198–205. doi:10.1177/1534508407313480

Clarke, B., Baker, S. K., 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

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.

Clements, D. H. (2003). Geometric and spatial thinking in early childhood education. In J. Sarama, D. H. Clements, & A.-M. DiBiase (Eds.), *Engaging Young Children in Mathematics: Standards for early childhood mathematics education* (pp. 267–298). Mahwah, NJ: Routledge Chapman & Hall.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.

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.

Dehaene, S. (2001). Précis of the number sense. *Mind & Language*, *16*(1), 16–36. doi:10.1111/1468-0017.00154

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

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(1), 128–141.

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

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

Hintze, J. M., Christ, T. J., & Keller, L. A. (2002). The generalizability of CBM survey-level mathematics assessments: Just how many samples do we need? *School Psychology Review*, *31*(4), 514–528.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi:10.1080/10705519909540118

Jenkins, J. R., Graff, J. J., & Miglioretti, D. L. (2009). Estimating reading growth using intermittent CBM progress monitoring. *Exceptional Children*, *75*(2), 151–163.

Jordan, N. C., Kaplan, D., Locuniak, M. N., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice*, *22*(1), 36–46. doi:10.1111/j.1540-5826.2007.00229.x

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

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., Deimann, P., & Kastner-Koller, U. (2002). DEMAT1+. Deutscher Mathematiktest für erste Klassen. [German mathematics test for first grades]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 34*(4), 236–238. doi:10.1026//0049-8637.34.4.238

Krajewski, K., Liehm, S., & Schneider, W. (2004). *DEMAT2+. 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

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

Muthén, L., & Muthén, B. O. (1998-2010). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

National Council of Teachers of Mathematics (2012). *Math standards and expectations*. Retrieved from http://www.nctm.org/standards/content.aspx?id=314

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23–74.

Shapiro, E. S. (2013). Commentary on progress monitoring with CBM-R and decision making: problems found and looking for solutions. *Journal of school psychology*, *51*(1), 59–66. doi:10.1016/j.jsp.2012.11.003

Souvignier, E., & Förster, N. (2011). Effekte prozessorientierter Diagnostik auf die Entwicklung der Lesekompetenz leseschwacher Viertklässler [Effects of curriculum-based measurement on reading achievement in fourth graders]. *Empirische Sonderpädagogik*, (3), 243–255.

Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research & Practice*, *15*(3), 128–134. doi:10.1207/SLDRP1503_2

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

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762. doi:10.1037/a0027627

# PART V

Appendices

# Curriculum Vitae

*- not part of the digital version of the dissertation -*

# Publications, conference presentations & reports

## Publications

Salaschek, M., & Souvignier, E. (accepted). Web-based progress monitoring in first grade mathematics. *Frontline Learning Research*.

Souvignier, E., Förster, N., & Salaschek, M. (in press). quop: ein Ansatz internet-basierter Lernverlaufsdiagnostik und Testkonzepte für Mathematik und Lesen [quop: an internet-based approach to learning progress assessment and test concepts for mathematics and reading]. In M. Hasselhorn, W. Schneider, & U. Trautwein (Eds.), *Formative Leistungsdiagnostik (Test und Trends N.F. Band 12)*. Göttingen, Germany: Hogrefe.

Behrmann, L., Förster, N., Salaschek, M., Schulte, E., & Souvignier, E. (2012). Verbesserung faktenbezogenen und konzeptuellen Wissens durch Lerntagebücher in der Hochschullehre [Improving fact-based and conceptual knowledge with learning diaries in university education]. In M. Krämer, S. Dutke, & J. Bahrenberg (Eds.), *Psychologiedidaktik und Evaluation IX* (pp. 293–300). Aachen, Germany: Shaker.

Förster, N., Behrmann, L., Salaschek, M., Schulte, E., & Souvignier, E. (2012). Lerntagebücher für alle? Für welche Studierenden stellen Lerntagebücher eine optimale Unterstützung dar? [Learning diaries for everyone? For which students are learning diaries an ideal aid?]. In M. Krämer, S. Dutke, & J. Bahrenberg (Eds.), *Psychologiedidaktik und Evaluation IX* (pp. 285–292). Aachen, Germany: Shaker.

Keuter, M., Salaschek, M., & Thielsch, M. T. (2012). Typologie der deutschen Onlinebevölkerung [Typology of the German online population]. In H. Reiterer & O. Deussen (Eds.), *Mensch & Computer 2012* (pp. 325–328). Munich, Germany: Oldenbourg.

Thielsch, M. T., Meese, C., & Salaschek, M. (2012). Datenschutz, Datensicherheit und ethische Aspekte in der Lehrevaluation [Data protection, data security and ethical aspects in course evaluations]. In M. Krämer, S. Dutke, & J. Bahrenberg (Eds.), *Psychologiedidaktik und Evaluation IX* (pp. 395–400). Aachen, Germany: Shaker.

Salaschek, M. (2009). Online user typology and aesthetics [Abstract]. In M. Welker, H. Geißler, L. Kaczmirek, & O. Wenzel (Eds.), *Proceedings of the 11th General Online Research Conference GOR 09* (p. 120). Hürth, Germany: German Society for Online-Research.

Salaschek, M., Holling, H., Freund, P. A., & Kuhn, J.-T. (2007). Benutzbarkeit von Software: Vor- und Nachteile verschiedener Methoden und Verfahren [Software usability: Pros and Cons of different methods and approaches]. *Zeitschrift für Evaluation*, *6*(2), 247–276.

Holling, H., Freund, P. A., Kuhn, J.-T., Salaschek, M., Gawlista, C., & Thielsch, M. T. (2006). Share your knowledge: Usability von Wissensmanagementsystemen [Share your knowledge: Usability of knowledge management systems]. In T. Bosenick, M. Hassenzahl, M. Müller-Prove, & M. Peissner (Eds.), *Usability Professionals 2006* (pp. 95–101). Stuttgart, Germany: German Chapter of the Usability Professionals Association.

## Conference presentations

Förster, N., Schulte, E., Salaschek, M., & Souvignier, E. (2013, August). *Assessment of reading progress in fourth grade: Differentiating between fluency and comprehension.* Paper presented at the 14th biennial meeting of the European Association of Research on Learning and Instruction (Earli), Munich, Germany.

Salaschek, M., Schulte, E., Förster, N., & Souvignier, E. (2013, August). *Predicting maths performance in primary school: Results of a computer-based progress monitoring tool.* Paper presented at the 14th biennial meeting of the European Association of Research on Learning and Instruction (Earli), Munich, Germany.

Schulte, E., Förster, N., Salaschek, M., & Souvignier, E. (2013, August). *Individualized reading instructions fosters reading fluency of three and four graders.* Poster presented at the 14th biennial meeting of the European Association of Research on Learning and Instruction (Earli), Munich, Germany.

Salaschek, M., Meese, C., & Thielsch, M. T. (2013, March). *Ethics and data security in web-based course evaluation*. Poster presented at the 15th annual meeting of General Online Research (GOR 13), Mannheim, Germany.

Förster, N., Behrmann, L., Salaschek, M., Schulte, E., & Souvignier, E. (2012, September). *„Ich kenne meine Schüler" – Diagnostische Kompetenz von Lehrkräften ["I know my students" – diagnostic competence of teachers].* Poster presented at the 48th annual meeting of Deutsche Gesellschaft für Psychologie (DGPs), Bielefeld, Germany.

Keuter, M., Salaschek, M., & Thielsch, M. T. (2012, September). *Typologie der deutschen Onlinebevölkerung [Typology of the German online population].* Poster presented at the annual Mensch & Computer meeting, Konstanz, Germany.

Salaschek, M., & Souvignier, E. (2012, July). *Online Progress monitoring in mathematics for second-graders*. Paper presented at the 10th biennial meeting of Junior Researchers of the European Association of Research on Learning and Instruction (JURE), Regensburg, Germany.

Behrmann, L., Förster, N., Salaschek, M., Schulte, E., & Souvignier, E. (2012, May). *Verbesserung faktenbezogenen und konzeptuellen Wissens durch Lerntagebücher in der Hochschullehre [Improving fact-based and conceptual knowledge with learning diaries in university education].* Paper presented at the 9th annual meeting of Psychologiedidaktik und Evaluation, Münster, Germany.

Förster, N., Behrmann, L., Salaschek, M., Schulte, E., & Souvignier, E. (2012, May). *Lerntagebücher für alle? Für welche Studierenden stellen Lerntagebücher eine optimale Unterstützung dar? [Learning diaries for everyone? For which students are learning diaries an ideal aid?].* Poster presented at the 9th annual meeting of Psychologiedidaktik und Evaluation, Münster, Germany.

Thielsch, M. T., Meese, C., & Salaschek, M. (2012, May). *Datenschutz, Datensicherheit und ethische Aspekte in der Lehrevaluation [Data protection, data security and ethical aspects in course evaluations].* Poster presented at the 9th annual meeting of Psychologiedidaktik und Evaluation, Münster, Germany.

Salaschek, M., & Souvignier, E. (2012, April). *Online curriculum-based measurement for second-graders in mathematics*. Paper presented at the 2012 annual meeting of the American Educational Research Association (AERA), Vancouver, BC, Canada.

Salaschek, M., Förster, N., & Souvignier, E. (2011, September). *Online curriculum-based measurement for first-graders in mathematics.* Poster presented at the 13th biennial meeting of the European Association of Research on Learning and Instruction (Earli), Exeter, United Kingdom.

Salaschek, M., & Souvignier, E. (2011, September). *Prognostische Validität eines internetgestützten Verfahrens zur frühen Lernverlaufsdiagnostik im Fach Mathematik [Prognostic validity of an internet-based approach to early learning progress assessment in mathematics].* Paper presented at the 13th annual meeting of Pädagogische Psychologie der DGPS, Erfurt, Germany.

Salaschek, M. (2009, April). *Online user typology and aesthetics*. Paper presented at the 11th annual meeting of General Online Research GOR 09, Vienna, Austria.

Holling, H., Freund, P. A., Kuhn, J.-T., Salaschek, M., Gawlista, C., & Thielsch, M. T. (2006, August). *Share your knowledge: Usability von Wissensmanagementsystemen [Share your knowledge: Usability of knowledge management systems].* Paper presented at the annual meeting of the German Chapter Usability Professionals' Association (UPA), Gelsenkirchen, Germany.

## Reports

Dusend, C., Salaschek, M., Thielsch, M. T., Stegemöller, I., & Fischer, S. (2012). *Evaluationsbericht Psychologie 2012: Gemeinsamer Bericht über die Evaluationen im Fach Psychologie im WiSe 11/12 und SoSe 12 [Evaluations report 2012: joint report on the evaluations in psychology in the fall semester 11/12 and the spring semester 11].* University of Münster, Germany.

Thielsch, M. T., Salaschek, M., Dusend, T., Grötemeier, I., & Fischer, S. (2011). *Evaluationsbericht Psychologie 2011: Gemeinsamer Bericht über die Evaluationen im Fach Psychologie im WiSe 10/11 und SoSe 11 [Evaluations report 2011: joint report on the evaluations in psychology in the fall semester 10/11 and the spring semester 11].* University of Münster, Germany.

Thielsch, M. T., Hirschfeld, G., Salaschek, M., Dusend, T., Grötemeier, I., & Fischer, S. (2010). *Evaluationsbericht Psychologie 2010: Gemeinsamer Bericht über die Evaluationen im Fach Psychologie im WiSe 09/10 und SoSe 10 [Evaluations report 2010: joint report on the evaluations in psychology in the fall semester 09/10 and the spring semester 10].* University of Münster, Germany.

Thielsch, M. T., Förster, N., Hirschfeld, G., Salaschek, M., & Hüttemann, T. (2008). Diagnostik-Online III: E-Learning in der psychologischen Diagnostikausbildung. [Diagnostics online III: e-learning in the teaching of psychological diagnostics]. In H. L. Grob (Ed.), *E-Learning Praxisberichte*, Münster, Germany.

Holling, H., Freund, P. A., Kuhn, J.-T., & Salaschek, M. (2006). *Benutzbarkeit von Software: Wie usable sind Evaluations-Verfahren? [Usability of software: How usable are evaluation approaches?].* Münster, Germany: ERCIS.

# Erklärung des Promovenden / der Promovendin zum eigenen Anteil an den vorgelegten wissenschaftlichen Abhandlungen mit zwei oder mehr Autor(inn)en (kumulative Dissertation)

| | |
|---|---|
| Promovend/Promovendin: | **Martin Salaschek** |
| Titel der Dissertation: | **Mathematics Progress Monitoring in the Primary Grades: Construction and Validation of Progress Monitoring Test Concepts in Grade 1 and 2** |

## Wissenschaftliche Abhandlung 1

| | |
|---|---|
| *Titel:* | Web-based progress monitoring in first grade mathematics |
| *Autor(en):* | Martin Salaschek & Elmar Souvignier |
| *Journal:* | Frontline Learning Research |
| *Publikationsstatus:* | □ nicht eingereicht<br>□ eingereicht<br>□ in Begutachtung<br>□ in Revision<br>☒ angenommen<br>□ veröffentlicht *Publikationsjahr:* |

*Beschreibung des eigenen Anteils, wenn **keine** Alleinautorenschaft vorliegt:*

Die Konzeption, Identifizierung des wissenschaftlichen Problems sowie die Entwicklung des Untersuchungsdesigns erfolgte in enger Kooperation mit dem Koautor. Angelehnt an Vorarbeiten des Koautors wurde das Untersuchungsmaterial im Wesentlichen von mir entwickelt. Die Datenerhebung, Aufbereitung und Auswertung der Daten wurde von mir übernommen. Die Interpretation und Diskussion der Daten erfolgte in Zusammenarbeit mit dem Koautor.

Die Verschriftlichung der Arbeit wurde im Wesentlichen von mir vorgenommen, Revisionen erfolgten in Absprache mit dem Koautor.

## Wissenschaftliche Abhandlung 2

| | |
|---|---|
| *Titel:* | Mathematics growth trajectories in first grade: Cumulative vs. compensatory patterns and the role of number sense |
| *Autor(en):* | Martin Salaschek, Nina Zeuch & Elmar Souvignier |
| *Journal:* | Learning and Individual Differences |
| *Publikationsstatus:* | □ nicht eingereicht<br>□ eingereicht<br>☒ in Begutachtung<br>□ in Revision<br>□ angenommen<br>□ veröffentlicht *Publikationsjahr:* |

*Beschreibung des eigenen Anteils, wenn **keine** Alleinautorenschaft vorliegt:*

Die Konzeption, Identifizierung des wissenschaftlichen Problems sowie die Entwicklung des Untersuchungsdesigns erfolgte in enger Kooperation mit dem Drittautor. Angelehnt an Vorarbeiten des Drittautors wurde das Untersuchungsmaterial im Wesentlichen von mir entwickelt. Die Datenerhebung und Aufbereitung der Daten wurde von mir übernommen. Teile der Auswertung (nämlich LGCM- und LCGA-Analysen) wurden im Wesentlichen von der Zweitautorin durchgeführt. Die Interpretation und Diskussion der Daten erfolgte in Zusammenarbeit der Zweitautorin und dem Drittautor.

Die Verschriftlichung der Arbeit wurde im Wesentlichen von mir vorgenommen, Revisionen erfolgten in Absprache mit der Zweitautorin und dem Drittautor.

## Wissenschaftliche Abhandlung 3

| | |
|---|---|
| *Titel:* | Web-based mathematics progress monitoring in second grade |
| *Autor(en):* | Martin Salaschek & Elmar Souvignier |
| *Journal:* | Journal of Psychoeducational Assessment |
| *Publikationsstatus:* | □ nicht eingereicht<br>□ eingereicht<br>☒ in Begutachtung<br>□ in Revision<br>□ angenommen<br>□ veröffentlicht *Publikationsjahr:* |

*Beschreibung des eigenen Anteils, wenn **keine** Alleinautorenschaft vorliegt:*

Die Konzeption, Identifizierung des wissenschaftlichen Problems sowie die Entwicklung des Untersuchungsdesigns erfolgte in enger Kooperation mit dem Koautor. Angelehnt an eigene Vorarbeiten und an Vorarbeiten des Koautors wurde das Untersuchungsmaterial im Wesentlichen von mir entwickelt. Die Datenerhebung, Aufbereitung und Auswertung der Daten wurde von mir übernommen. Die Interpretation und Diskussion der Daten erfolgte in enger Zusammenarbeit mit dem Koautor.

Die Verschriftlichung der Arbeit wurde im Wesentlichen von mir vorgenommen, Revisionen erfolgten in Absprache mit dem Koautor.

*Bitte Seite 2 mehrmals ausfüllen, falls die Dissertation aus mehr als 3 wissenschaftlichen Abhandlungen besteht.*

Ort, Datum

Unterschrift Promovend(in)

**Zur Erläuterung: Beschreibung des Eigenanteils**
Bei der Angabe Ihrer Eigenanteile an Abhandlungen mit mehreren Autoren können Sie sich an den gängigen Kriterien internationaler "peer-reviewed" Fachzeitschriften orientieren:

1. Worin besteht ihr eigener intellektueller Anteil an dieser Studie? Dies können Anteile u.a. an der Konzeption, der Identifizierung des wissenschaftlichen Problems, der Entwicklung des Untersuchungsdesigns, der Erstellung des Untersuchungsmaterials, der Aufbereitung, Auswertung, Interpretation und Diskussion der Daten sein.
2. Was ist Ihr Anteil an der Verschriftlichung der Arbeit, also bei der Abfassung des Manuskripts selbst?