Studies in Educational Evaluation 59 (2018) 133–140

Contents lists available at ScienceDirect

Studies in Educational Evaluation

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

Does the textbook matter? Longitudinal effects of textbook choice on primary school students' achievement in mathematics

Ann-Katrin van den Ham*, Aiso Heinze

Leibniz Institute for Science and Mathematics Education (IPN), Olshausenstr. 62, 24118 Kiel, Germany

ARTICLE INFO

Keywords:
Textbook effects
Development of mathematics achievement
Program evaluation
School-based evaluation

ABSTRACT

Mathematics textbooks are ascribed an important role for classroom practice. Until now there are still open questions concerning the genuine effect of textbooks on students' learning in mathematics. This paper examines the effect of different textbooks representing the same curriculum on the student achievement by reanalyzing a longitudinal data set on primary school students' mathematics skills from Grade 1 to 3 (N = 1664). Results from multilevel regression analyses showed that mathematics teachers' textbook choice has a substantial effect on the students' mathematics achievement and that individual textbooks substantially differ in their effects. Furthermore, there are indications that the effect of textbook choice is cumulative over the school years. The findings suggest that textbooks should be considered as an important covariate in educational research and that textbook choice is a relevant factor for educational practice.

1. Introduction

Mathematics textbooks have an important role in representing and translating the abstract curriculum into operations that teachers and students can carry out (Valverde, Bianchi, Wolfe, Schmidt, & Houang, 2002). They are extensively used in everyday classroom practice (Mullis, Martin, Foy, & Arora, 2012), differ in their content and pedagogical styles (Pepin & Haggarty, 2001) and therefore shape the potential learning opportunities for students. Hence, it is commonly assumed that textbooks have a substantial effect on student achievement. Presently, there is a lack of empirical evidence supporting this assumption. Research on textbook effects reveals conflicting results (van Steenbrugge, Valcke, & Desoete, 2013), and is often based on small sample sizes and cross-sectional designs (Fan, Zhu, & Miao, 2013). Many textbook studies – especially those with larger sample sizes – are in fact curriculum studies because the considered textbooks represent different curricula (e.g., standard based curriculum vs. traditional curriculum, Tarr et al., 2008; Koedel, Morgan, Polikoff, Hardaway, & Wrabel, 2017). The aim of the present study is therefore to examine the effect of different textbooks on student achievement, representing the same curriculum. The results are based on a reanalysis of a longitudinal data set on primary school students' mathematics skills from Grade 1 to 3.

1.1. Role of the textbook

Textbooks are artifacts since they are educational material created by human beings (e.g., Rabardel, 2002). They are written by an author or group of authors and produced by a publisher. The authors interpret a curriculum and transform it into learning opportunities and concrete operations that teachers and students can carry out (Valverde et al., 2002). Hence, a textbook can be described as a mediator between the intended curriculum as official policy and the implemented curriculum by the teachers (Valverde et al., 2002). Textbooks are therefore seen as conveyors of the curriculum and often referred to as curriculum material (e.g. Herbel-Eisenmann, 2007; Remillard, 2005). They offer teachers and students opportunities for teaching and learning by providing an objectively given didactical structure of the content. Textbooks thereby specify a certain manner of use and therefore limit possible uses (Remillard, 2005; Rezat, 2008).

Due to their mediating role between intended and implemented curriculum, textbooks can be used as a monitoring instrument in the educational system. In some countries textbooks must be approved and licensed by the ministry of education (e.g., Hong Kong, Norway), while in other countries there is no approving authority (e.g., The Netherlands, some states in the US) (Mullis et al., 1997). The strictness of monitoring by an authority can influence the curriculum interpretation in the textbook (Schmidt, Raizen, Britton, Bianchi, & Wolfe, 1997). Summarizing, mathematics textbooks are ascribed an important

* Corresponding author.
E-mail address: vandenham@ipn.uni-kiel.de (A.-K. van den Ham).

https://doi.org/10.1016/j.stueduc.2018.07.005
Received 9 April 2018; Received in revised form 12 July 2018; Accepted 16 July 2018
0191-491X/

role for classroom practice (Pepin & Haggarty, 2001). Even though one can assume that teachers and students modify the offered learning opportunities in textbooks, it is assumed that textbooks exert influence on the teaching and learning in the classroom (Valverde et al., 2002).

1.2. Textbook content

Textbooks are artefacts and therefore historically developed, culturally formed, produced for certain ends and used with particular intentions (Rezat, 2008). Research shows that mathematics textbooks differ in their mathematical content as well as in certain pedagogical, cultural and sociological aspects (e.g., related to gender, ethnicity, equity, Fan et al., 2013). For example, cross-national comparisons revealed that mathematics textbooks vary across different countries or states due to specific cultural and educational traditions (Mayer, Sims, & Tajika, 1995; Pepin & Haggarty, 2001). Studies regarding the mathematical content indicate variations between textbooks concerning the interpretation of a curriculum (Fan et al., 2013). Differences occur, for example, with respect to the structure of the presented mathematical topic (e.g., Jones et al., 2015, for the statistical content in 25 primary school textbooks in Texas) and the cognitive demands of the learning opportunities (e.g., Kolovou, van den Heuvel-Panhuizen, & Bakker, 2009, for cognitive demands of number problems in the six main textbook series used in Dutch primary schools). To conclude, textbook differences are based on different curriculum interpretations following different pedagogical intentions, cultural or educational traditions. This generates different conditions for students' mathematics learning, depending on the teachers' use of textbooks for lesson preparation.

1.3. Use of the textbook in teaching and learning

Several studies revealed that teachers frequently use mathematics textbooks as the main basis for their instruction. For example, according to the Trends in International Mathematics and Science Study (TIMSS) 2011 on average about 75% of the primary school teachers base their instruction on the mathematics textbook. In Germany 86% of the teachers report using the mathematics textbook as a basis of instruction (Mullis, Martin, Foy, & Drucker, 2012). Lepik, Grevholm, and Viholainen, (2015) analyzed surveys from 402 mathematics teachers of Grade 7 to 9 from Estonia, Finland and Norway. Depending on the country, 49–64% of the teachers rely heavily on the textbook in terms of planning and preparing their lessons. Furthermore 79–92% of the teachers use the textbook as the only source for exercises in at least half of the lessons. Krammer (1985) analyzed data from systematic lesson observations of 50 mathematics teachers from 17 schools in the Netherlands which used one of the three best known Dutch mathematics textbooks. The results indicated that users of different textbooks implement different teaching practices. Evidence for a consistency of the relationship between textbook features and teaching practices is also provided by data from TIMSS 1995 and the related video study. The analyses of the United States TIMSS eighth grade data revealed a positive relation between the space a topic covers in a textbook and the instructional time teachers using this textbook have dedicated to this topic in the mathematics classroom (Schmidt et al., 2001). Mathematics textbooks influence what topics are covered and how these topics are presented. It was concluded that mathematical topics covered in given curriculum materials are considered of fundamental importance by the teachers (Stein, Remillard, & Smith, 2007), whereas there is a low probability that topics not covered in the textbook will be presented in the classroom (Schmidt et al., 1997).

The mathematics textbook is not only an important resource for teachers, but also for students. Although curriculum guides usually define the official curriculum for students, many students are not even aware of them (Schmidt et al., 1997). Textbooks on the other hand are a basis for everyday school practice (Mullis, Martin, Foy, Arora et al., 2012; Mullis, Martin, Foy, Drucker et al., 2012; Schmidt et al., 2001). In their study Lepik et al. (2015) found that most problems for students' in-class exercises and homework were taken from mathematics textbooks. Furthermore, 45% (Norway) to 76% (Finland) of the teachers used the textbook as the only source for homework. According to TIMSS 1995 students of most western countries report that they also work from textbooks or work cards on their own in most lessons (e.g., England 49%, United States 57%, New Zeeland 50%).

In summary it can be stated, that mathematics textbooks are an important and extensively used resource for teaching and student learning.

1.4. Influence of textbook choice on students' mathematics achievement

Schmidt et al. (2001) analyzed the United States TIMSS 1995 eighth grade data and found a direct relation between the amount of space allocated to covering a topic and the size of students' achievement gain on that topic. Similar findings regarding the interaction between learning opportunities in mathematics textbooks and learning outcomes were reported by Törnroos (2005). Törnroos examined the influence of nine mathematics textbooks series used in the classes of the TIMSS 1999 sample in Finland on students' achievement in the TIMSS mathematics test. It turned out that the number of learning opportunities a textbook provided specifically for the content of TIMSS items was significantly positively correlated with students' performance in the TIMSS test. Furthermore Hadar (2017) analyzed whether opportunities to engage in tasks demanding different levels of understanding provided in mathematics textbooks correlate with students' achievements on tasks demanding equivalent levels of understanding on a standardized national exam. She examined two 8th grade mathematics textbooks used by students in the Arab community in Israel. Using a sample of all 8th grade students in the Arab community who completed the national math test in 2015 and studied in schools using one of the two textbooks ($N = 4040$ students) Hadar found that students using a book will have higher scores if this textbook provides the opportunity to engage in tasks demanding higher levels of understanding. In contrast to this results, a study of van Steenbrugge et al. (2013) in Flanders (Belgium) did not find substantial differential effects of mathematics textbooks on students' achievement. The cross-sectional study included 1579 students (Grade 1–6) and their 89 teachers using five different mathematics textbook series. The authors conclude that "Up to date there is no agreement about the differential impact of mathematics textbooks on students' performance results" (van Steenbrugge et al., 2013, p. 346).

The previously presented research indicates that there are still open questions concerning the genuine effect of textbooks on students' learning in mathematics. Many research studies have limitations due to small sample sizes and/or cross-sectional designs (cf. the review article Fan et al., 2013). In the US, textbook effects were analyzed with adequate sample sizes in the context of curriculum research (e.g., Tarr et al., 2008; Koedel et al., 2017). However, the textbooks in these studies served as indicators for the implementation of different curricula and, consequently, the authors interpreted the effects on students' mathematics achievement as curriculum effects. Fan et al. (2013) concluded in their review of research focusing on mathematics textbooks that there is still a strong need for more confirmatory research about the relationship of mathematics textbooks and student achievement. Five years later to the best of our knowledge, there are no large longitudinal studies examining genuine textbook effects (i.e., effects of textbooks representing the same curriculum) on students' mathematics achievement.

1.5. The present study

We analyzed data of a three-year longitudinal study in primary school comprising teacher information as well as student information including arithmetic skills from Grade 1 and 2 as well as the students' scores of a nationwide mathematics competence test at the end of Grade 3. The sample was selected in one federal state of Germany (Schleswig-Holstein). As a result the textbooks in this sample represented the same statewide curriculum. The majority of the classes in the sample used one of four common mathematics textbooks which were distributed relatively evenly across the classes. Therefore, this sample allows us to examine the effects of four different primary school textbooks representing the same curriculum on students' mathematics achievement with a longitudinal design and a sound sample size. Accordingly, our study addresses the following research question:

Does the textbook choice of primary mathematics teachers have an effect on the development of students' mathematics achievement through Grade 1, 2 and 3?

Following this research question we developed three hypotheses:

(1) Given that teachers use textbooks as a basis for their instruction (e.g., Mullis, Martin, Foy, Arora et al., 2012; Mullis, Martin, Foy, Drucker et al., 2012; Lepik et al., 2015) and that textbooks influence the teaching and topics covered in a classroom (e.g., Schmidt et al., 1997, 2001; Stein et al., 2007), our first hypothesis is that the textbook choice has an effect on the student achievement even when the different textbooks are aligned to the same intended curriculum.
(2) As the textbooks vary in their content presentation (e.g., Fan et al., 2013; Kolovou et al., 2009; Jones et al., 2015), our second hypothesis is that textbooks aligned to the same intended curriculum differ in their effects on student achievement.
(3) Following the second hypothesis and the fact that mathematics learning is a cumulative process, our third hypothesis is that the textbook choice has a genuine effect on student achievement in each school year and the overall effect increases over the years.

2. Methods

2.1. Participants and design

Our study is a reanalysis of data that stems from a longitudinal evaluation of a mathematics support program for weak students in the regular mathematics classroom. In the evaluation a control group is compared with two groups in which teachers received additional teaching material for weak students in Grade 1 and 2. The participation in the support program is independent of the textbooks the teachers used and we controlled for the participation in the support program by dummy coded variables in the statistical analyses (see Section 2.4). The sample comprises 40 primary schools from urban and rural areas in the German federal state of Schleswig-Holstein. In this federal state all teachers have to follow the same state-wide curriculum for each grade which prescribes content (e.g., concepts, algorithms), skills (e.g., mental computation, estimation), and representations (e.g., number line, place value chart) to be addressed in the mathematics classroom.

The longitudinal sample consists of 2330 children from 127 classes from the 40 schools which is about 10% of the cohort in this federal state. About three-quarters of the classes in the sample used one of four mathematics textbook series. The distribution of these 93 classes over the textbook series is quite uniform (see Section 3.1). Therefore, this subsample of 93 classes with 1664 children[1] (784 female, 47%) from 38 schools is a sound basis for examining our research question.

The four textbooks series in our sample show specific characteristics. The textbook series "Denken und Rechnen" and "Welt der Zahl" provide a usual textbook format (i.e., one book for each grade covering the content for one school year) whereas "Flex und Flo" and "Einstern" provide a set of 4–7 booklets for each grade separating specific topics. For each grade, series "Denken und Rechnen", "Flex und Flo" and "Welt der Zahl" are comparable in their approach by offering teachers freedom to select and use the examples, tasks, representations, explanations, and exercises provided in the textbook; a specific learning trajectory for students is suggested but not prescribed. In contrast, the textbook series "Einstern" recommends using the booklets and working through each booklet in a linear order. This series especially fosters students' individual and autonomous learning. The content is structured in small parts and typically each aspect has 1–2 pages dedicated to exercises.

2.2. Measures

2.2.1. Mathematics achievement

Students' learning progress at the end of Grade 1 and 2 was measured with curriculum based arithmetic tests. The arithmetic content domain holds the biggest proportion of the German primary school curriculum. The items for these tests were adapted from the approved tests of the "personality and learning development of primary school children" study (Persönlichkeits- und Lernentwicklung von Grundschulkindern, PERLE) from Lipowsky et al. (2011). Among other things, the test at the end of Grade 1 (39 items, 30 min) addressed doubling, halving, part-whole relations, addition, and subtraction in the number domain 0–20, whereas the test at the end of Grade 2 (59 items, 40 min) covered, for example, addition, subtraction, multiplication, division in the number domain 0–100, doubling, halving, factors of a given number. The short-constructed response format was used (response in form of a number). Some of the items were combined and in those cases partial credit coding was used. Both tests were linked by 19 common items so that the same scale could be used for both tests (see Section 2.4). The reliability of the adapted tests were very good (EAP/PV reliability = .93 and .94). The effect size (Cohen's d) for the achievement growth from the end of Grade 1 to the end of Grade 2 was $d = 1.06$.

Regarding students' mathematics achievement at the end of Grade 3, the total scores from the national mathematics test for the two content areas "numbers" and "patterns" were provided by the schools. The national mathematics test is intended to measure to what extend the national educational standards are met at the end of Grade 3 (cf. Zimmer-Müller, 2013). The items of the domain numbers address all four basic operations in the number domain 0–1000, representations of numbers as well as number relations. The domain patterns contains items regarding understanding and using functional relations between mathematical objects and patterns. The character of the items in both domains differ from those of our arithmetic tests in Grade 1 and 2. Following the idea of the national educational standards, the items mainly address the application of arithmetical knowledge in mathematical and realistic contexts.

2.2.2. Control measures

Data about the textbooks, textbook use and teacher qualification were collected using a teacher questionnaire. In order to control individual differences between individual students as well as classes, students' learning prerequisites related to basic numerical skills and German language skills were measured at school entrance using approved standardized tests (basic numerical skills: Lorenz, 2007, Cronbach's $\alpha = .74$; German language skills: Münsteraner Screening, MÜSC, Mannhaupt, 2013, Cronbach's $\alpha = .72$). Students' basic cognitive abilities were measured using the Culture Fair Intelligence test (CFT 1-R, Weiß & Osterland, 2013, Cronbach's $\alpha = .91$).

2.3. Procedure

Data collection started in the school year 2013/2014. At the beginning of Grade 1, learning prerequisites related to basic numerical skills and language skills and halfway through Grade 1 cognitive

[1] For students' achievement at the end of grade 3 only total test scores for 1628 students were available in the data set.

abilities were assessed. At the end of Grade 1 and 2 data for arithmetic achievement was collected. All tests were administered by trained test administrators according to the respective manuals.

2.4. Data analyses

Because of the anchor item design of the arithmetic test in Grade 1 and 2 Item Response Theory was used to estimate the student scores in Grade 1 and 2 on the same scale. Since the tests contained partial credit items an extension of the Rasch model, the partial credit model (PCM; Masters, 1982), was used. The student scores were estimated in a two-step procedure. First, the tests were calibrated, using the data of all students who attended the respective test to monitor the quality of the items and estimate the item parameters. Second, five plausible values for each of the 1664 students were generated by fitting a latent regression model and using item parameters anchored at their estimated values from the previous calibration[2]. A detailed review of the plausible values methodology is given in Mislevy (1991). For the latent regression model, additional variables from the teacher and the parent questionnaire were aggregated to orthogonal factors and used as background information. This method was chosen because it allows for unbiased population level analyses of competence distributions and context variables (von Davier, Gonzalez, & Mislevy, 2009).

To analyze textbook effects we used multilevel random intercept models to control for the nested structure of the data (students nested within classes)[3]. First, models without predictors (null models) were estimated to determine the partition of variance between and within classes. Second, models were estimated including individual characteristics (at individual level) and class composition (at class level). The learning prerequisites were included on individual level to account for the value added in the school period and to account for individual differences between pupils. For all three variables the individual scores were aggregated to class means and included on class level to control for class composition. Since the mathematics support programs were intended to have a positive effect on student achievement and since in the federal state of Schleswig-Holstein about 40% of the mathematics primary school teachers have no formal qualification for teaching mathematics, we controlled for these variables in a third model. Therefore, the support programs and teacher qualification (studied mathematics or not) were included as dummy coded variables. In the fourth model, we included the textbook choice as dummy coded variables on the class level (the textbook Einstern was specified as reference category). Finally, for the effects at the end of Grade 2 and 3 we included in a fifth model the mathematics achievement of the previous grade (on the individual level and as class mean on the class level) to investigate the effect of textbook choice on the achievement gain during the respective school year.

Scores for the learning prerequisites as well as the mathematics achievement scores were standardized so that the corresponding ß-coefficients can be interpreted as effect sizes similar to Cohen's *d* (cf. Tymms, 2004). For missing data at the independent variables for the learning prerequisites we applied a full-information-maximum-likelihood (FIML) approach in Mplus 7.0 (Muthén & Muthén, 1998-2012Muthén & Muthén, 1998-2012). FIML combines missing data and parameter estimation in a single step and uses all the available information (Enders, 2010). The percentage of missing data in the learning prerequisites was between 6.6% (CFT) and 8.8% (MÜSC). The tests were a non-obligatory part of the normal mathematics lesson and we did not detect specific patterns in the missingness.

Due to sample selection, there was no missing data for the predictors textbook choice and support program.

3. Results

First, we present a short summary of descriptive information about the teacher qualification, the textbooks and their use as well as the learning prerequisites.

3.1. Descriptive results

About half of the 93 teachers in our sample (51 teachers, 54.8%) had a formal qualification for teaching mathematics, whereas the other teachers were qualified for other subjects and teach mathematics as "generalists". The four textbooks used in our sample were distributed relatively evenly, that is 18 teachers (19.4%) used the textbook "Denken und Rechnen", 21 (22.6%) "Welt der Zahl" and 21 (22.6%) "Einstern", while slightly more teachers (33, 35.5%) used the textbook "Flex und Flo".

The learning prerequisites and the dependent variables (mathematics achievement) in the four textbook groups are reported in Table 1. To ensure that the teachers of this sample extensively use the textbooks, they were asked how often they use the textbook when teaching arithmetic. Of the 93 teachers 87 replied and the vast majority of them (84 teachers, 96.6%) reported that they use the textbook (almost) every day (68 teachers, 78.2%) or one to two times a week (16, 18.4%). Furthermore the teachers were asked to what extent they agreed with the statement that the textbook guided their teaching regarding content, method and materials they used. Here, 86 teachers replied and 75 of these (87.2%) agreed or strongly agreed with the statement. These findings imply that the textbooks in our sample can be considered as a main source for the lesson preparation and mathematics instruction.

We tested whether teacher qualification and textbook use is independent from the textbook choice. The number of teachers having a formal qualification did not differ significantly between the textbooks ($\chi^2(3, N = 92) = 5.58$, $p = .13$). The four textbook groups did not differ significantly in the frequency of textbook use ($H(3) = 2.24$, $p = .52$) and in the reliance on the textbook ($H(3) = 0.94$, $p = .82$)[4].

Null models were estimated in the multilevel analyses to determine the share of between-class variance (i.e., given by the intra-class coefficient, ICC) in students' mathematics achievement. In these models, 14% (arithmetic Grade 1), 19% (arithmetic Grade 2) and 15% respectively 16% (numbers/patterns Grade 3) of the variance laid between classes. Hence, the mathematics achievement differed depending on the class attended by the student.

3.2. Results of the multilevel analyses: control measures

Regarding individual characteristics at school entrance (see Models 1 in Tables 2–4) cognitive abilities and basic numerical skills showed a significant effect on the mathematics achievement at the end of Grade 1, 2 and 3, whereas language skills had a significant effect only at the end of Grade 1. The individual learning prerequisites explained 40.5% of the variance within classes at the end of Grade 1 and decrease to 33.6% at the end of Grade 2 and 33.6% (numbers) respective 34.5% (patterns) at the end of Grade 3.

The class composition regarding the learning prerequisites in numerical and language skills showed no significant differential effect on the mathematics achievement at the end of Grade 1, 2 and 3. The composition regarding the cognitive abilities yielded relatively large

[2] Of the 1664 students 16.5% didn't take the test at the end of Grade 2. Using the described method for estimating plausible values the student abilities of those students were imputed in the process.

[3] The assumptions of the multilevel regression, in particular, homoscedasticity, normality, and the absence of multicollinearity, were met.

[4] The Kruskal-Wallis test was used because the independent variable was not distributed normally and the expected frequencies in seven categories were smaller than five.

Table 1
Mean values and standard deviations for individual learning prerequisites and mathematics achievement in Grade 1–3 (dependent variables) in the four textbook groups.

Textbook choice	Learning prerequisites at school entrance			Mathematics achievement			
	Cognitive abilities	Basic numerical skills	Language skills	Arithmetic achievement Grade 1	Arithmetic achievement Grade 2	Achievement Grade 3 (Numbers)	Achievement Grade 3 (Patterns)
Denken und Rechnen	26.25 (8.63)	30.93 (7.40)	32.65 (6.14)	−0.09 (1.03)	0.21 (1.00)	11.07 (3.34)	8.00 (2.62)
Einstern	25.46 (8.58)	32.05 (6.34)	33.42 (5.80)	−0.17 (0.96)	−0.29 (0.91)	9.47 (3.44)	6.83 (2.66)
Flex und Flo	27.27 (9.23)	31.10 (7.86)	32.96 (6.37)	−0.02 (0.98)	0.01 (1.02)	10.41 (3.33)	7.32 (2.61)
Welt der Zahl	28.41 (9.03)	32.40 (6.55)	34.51 (5.39)	0.28 (0.95)	0.23 (0.94)	11.37 (3.35)	8.21 (2.45)

Note. Arithmetic achievement at the end of Grade 1 and 2 are z-values computed from the IRT model whereas all other results are sum scores from the tests. The maximum score for cognitive abilities is 44, for basic numerical skills 43, for language skills 40 and for achievement Grade 3 (Numbers/Patterns) 12/17.

Table 2
Multilevel regression for individual and classroom covariates and textbook on students' arithmetic achievement at the end of Grade 1.

	Model 1	Model 2	Model 3
Individual level			
Cognitive abilities	.29** (.03)	.29** (.03)	.29** (.03)
Basic numerical skills	.28** (.04)	.28** (.04)	.28** (.04)
Language skills	.14** (.03)	.14** (.03)	.14** (.03)
Class level			
Cognitive abilities (mean)	.20 (.11)	.21 (.11)	.12 (.11)
Basic numerical skills (mean)	.03 (.11)	.05 (.11)	.11 (.10)
Language skills (mean)	.02 (.12)	−.01 (.11)	−.04 (.10)
Teacher qualification		−.07 (.06)	−.09 (.06)
Support program 1		−.14* (.07)	−.10 (.07)
Support program 2		−.10 (.09)	−.05 (.08)
Denken und Rechnen			.16 (.10)
Flex und Flo			.14 (.09)
Welt der Zahl			.28 ** (.08)
Intercept	.01 (.03)	.14* (.06)	−.03 (.08)
Explained within class variance (%)	40.5%	40.5%	40.5%
Explained between class variance (%)	13.8%	20.6%	32.3%

** $p < .01$.
* $p < .05$; standard errors are in parentheses, variables on individual level and the output variable are standardized, cognitive abilities, numerical and language skills on class level are arithmetic means, Support program 1 and 2 are dummy variables with reference category control group, textbooks "Denken und Rechnen", "Flex und Flo", "Welt der Zahl" are dummy variables with reference category textbook "Einstern".

Table 3
Multilevel regression for individual and classroom covariates and textbook on students' arithmetic achievement at the end of Grade 2.

	Model 1	Model 2	Model 3	Model 4
Individual level				
Cognitive abilities	.35** (.03)	.35** (.03)	.35** (.04)	.23** (.03)
Basic numerical skills	.20** (.04)	.20** (.04)	.20** (.04)	.06 (.19)
Language skills	.06 (.03)	.06 (.03)	.06 (.03)	−.01 (.03)
Arithmetic skills (Grade 1)				.47** (.03)
Class level				
Cognitive abilities (mean)	.39* (.16)	.41* (.16)	.30* (.14)	.24 (.13)
Basic numerical skills (mean)	.09 (.13)	.08 (.12)	.14 (.11)	.04 (.09)
Language skills (mean)	−.07 (.14)	−.08 (.14)	−.09(.13)	−.05 (.10)
Arithmetic skills (Grade 1, mean)				.11 (.09)
Teacher qualification		−.04 (.08)	−.07 (.07)	−.04 (.06)
Support program 1		.14 (.08)	.15* (.08)	.21** (.07)
Support program 2		.16 (.09)	.25** (.09)	.28** (.08)
Denken und Rechnen			.43** (.10)	.35** (.08)
Flex und Flo			.20* (.09)	.15 (.07)
Welt der Zahl			.33** (.10)	.20* (.09)
Intercept	.04 (.04)	−.06 (.07)	−.31** (.08)	−.30** (.07)
Explained within class variance (%)	33.6%	33.6%	33.6%	50.1 %
Explained between class variance (%)	26.4%	30.9%	51.7%	56.8 %

** $p < .01$.
* $p < .05$; standard errors are in parentheses, variables on individual level and the output variable are standardized, cognitive abilities, numerical and language skills on class level are arithmetic means, Support program 1 and 2 are dummy variables with reference category control group, textbooks "Denken und Rechnen", "Flex und Flo", "Welt der Zahl" are dummy variables with reference category textbook "Einstern".

regression coefficients but reached the level of statistical significance only for arithmetic achievement at the end of Grade 2. Overall, the class composition explained between 7.1% and 26.4% of the variance between classes. Regarding learning environment characteristics (see Models 2 in Tables 2–4) the teacher qualification had no significant influence on the mathematics achievement in all grades. This is in line with the outcomes of the German National Study 2016 at the end of Grade 4 (Rjosk, Hoffmann, Richter, Marx, & Gresch, 2017).

The two support programs did not show systematic effects in this sample. At the end of Grade 1 there was a significant influence of the support program 1 on the mathematics achievement which disappeared when the textbook choice was included into the model (Model 3 in Table 2). At the end of Grade 2, both support programs showed only a significant effect after including the textbook choice into the model (Model 3 in Table 3).

Altogether the inclusion of the learning environment variables led to a small increase of the explained variance in all grades (ΔR^2 between 3.2% and 6.8%).

3.3. Results of the multilevel analyses: effects of textbook choice

When including the dummy variables for the textbook choice into the multilevel models (Models 3 in Tables 2–4), the explained variance increased substantially (Grade 1: $\Delta R^2 = 11.7\%$, Grade 2: $\Delta R^2 = 20.8\%$, Grade 3 numbers: $\Delta R^2 = 19.4\%$, Grade 3 patterns: $\Delta R^2 = 23.3\%$). The significant regression coefficients of the different textbooks varied between $\beta = .20$ and $\beta = .48$. Due to the standardized scales, the β-values can be interpreted as effect sizes similar to Cohen's d (see Section 2.4, cf. Tymms, 2004). Because the effect size for arithmetic achievement growth at the beginning of primary school was $d = 1.06$ in the first school year (see Section 2.2.1), these effects are of considerable size. Hence, we can confirm Hypothesis 1. The textbook choice has an effect on the student achievement in arithmetic at the end of Grade 1 and 2 as well as on the student achievement in the content domains "number" and "pattern" at the end of Grade 3.

Regarding Hypothesis 2 (textbooks differ in their effects on student achievement), it turned out that already at the end of Grade 1 the textbook "Denken und Rechnen" has a significantly positive effect ($\beta = .28$) on students' arithmetic achievement in comparison to the

Table 4
Multilevel regression for individual and classroom covariates and textbook on students' achievement for the competence dimensions "Numbers" and "Patterns" at the end of Grade 3.

	Model 1		Model 2		Model 3		Model 4	
	Numbers	Patterns	Numbers	Patterns	Numbers	Patterns	Numbers	Patterns
Individual level								
Cognitive abilities	.37** (.04)	.43** (.04)	.37** (.04)	.43** (.04)	.37** (.04)	.43** (.04)	.16** (.03)	.23** (.03)
Basic numerical skills	.20** (.05)	.17** (.05)	.20** (.05)	.17** (.05)	.20** (.05)	.17** (.05)	.08* (.03)	.06 (.04)
Language skills	.09* (.04)	.06 (.04)	.09* (.04)	.06 (.04)	.09* (.04)	.06 (.04)	.07* (.03)	.03 (.03)
Arithmetic skills (Grade 2)							.63** (.03)	.58** (.03)
Class level								
Cognitive abilities (mean)	.12 (.21)	.27 (.16)	.16 (.21)	0.31 (.16)	.03 (.19)	.25 (.15)	−.10 (.14)	.15 (.13)
Basic numerical skills (mean)	.08 (.17)	.08 (.16)	.10 (.17)	0.10 (.17)	.12 (.16)	.08 (.16)	.04 (.13)	.01 (.14)
Language skills (mean)	.03 (.17)	.01 (.16)	−.03 (.18)	−0.06 (.16)	.02 (.17)	.01 (.16)	.07 (.14)	.06 (.14)
Arithmetic skills (Grade 2, mean)							−.11 (.10)	−.13 (.09)
Teacher qualification			−.04 (.03)	−0.04 (.03)	.00 (.10)	.03 (.10)	.03 (.08)	.05 (.08)
Support program 1			−.02 (.10)	−0.13 (.11)	.01 (.10)	−.09 (.10)	.00 (.08)	−.10 (.09)
Support program 2			.05 (.11)	−0.09 (.12)	.14 (.11)	.00 (.11)	.01 (.09)	−.11 (.10)
Denken und Rechnen					.48** (.13)	.41** (.13)	.21 (.13)	.17 (.13)
Flex und Flo					.17 (.11)	.03 (.10)	.08 (.10)	−.05 (.10)
Welt der Zahl					.36** (.12)	.22* (.11)	.22 (.11)	.10 (.11)
Intercept	−.16** (.05)	−.17** (.04)	−.14 (.08)	−.04 (.10)	−.44** (.11)	−.28* (.11)	−.27 (.09)	−.13 (.11)
Explained within class variance (%)	33.6%	34.5%	33.6%	34.5%	33.8%	34.6%	61.9%	59.8%
Explained between class variance (%)	7.1%	17.2%	10.3%	23.1%	33.6%	42.5%	19.4%	25.0%

** $p < .01$, * $p < .05$; standard errors are in parentheses, variables on individual level and the output variable are standardized, cognitive abilities, numerical and language skills on class level are arithmetic means, Support program 1 and 2 are dummy variables with reference category control group, textbooks "Denken und Rechnen", "Flex und Flo", "Welt der Zahl" are dummy variables with reference category textbook "Einstern".

textbook "Einstern". At the end of Grade 2, "Denken und Rechnen" (β = .43), "Flex und Flo" (β = .20) and "Welt der Zahl" (β = .33) showed significantly positive effects in comparison to "Einstern". Regarding the scores of the national mathematics test at the end of Grade 3, there were significantly positive effects of "Denken und Rechnen" (numbers: β = .48/patterns: β = .41) and "Welt der Zahl" (numbers: β = .36/patterns: β = .22) in comparison to "Einstern". Comparing the textbooks "Denken und Rechnen", "Flex und Flo" and "Welt der Zahl", the Wald Test indicates that the effect of "Flex und Flo" was significantly lower than the effect of "Denken und Rechnen" at the end of Grade 2 and 3. Choosing "Flex und Flo" as reference category in the multilevel model, we got a significantly positive effect of "Denken und Rechnen" in comparison to "Flex und Flo" (Grade 2: β = .23, Grade 3 numbers: β = .31/patterns: β = .38). Hypothesis 2 can therefore be confirmed.

To examine Hypotheses 3 (effect of textbook choice increases during the first primary school years), we first compared the gain of explained variance between the Models 2 and 3 in each of the Tables 2–4. It turned out that the percentage of explained variance caused by the textbook variables increased from the end of the first grade to the end of Grades 2 and 3. This indicates that an increasing amount of variance in the student achievement can be ascribed to the textbook choice over the years. In a second step, we controlled for mathematics achievement of the previous school year to identify the genuine effect for the textbook choice in Grade 2 and 3 respectively. For Grade 2 (see Model 4 in Table 3), we still found significant effects for "Denken und Rechnen" (β = .35) and "Welt der Zahl" (β = .20) in comparison to the reference "Einstern", whereas the effect of "Flex und Flo" (β = .15) no longer reached the level of significance. For Grade 3 (see Model 4 in Table 4), we could not identify significant differences between the textbooks, which might be caused by the different test content in this grade. Overall, there are indications for cumulative effects of textbooks on students' achievement at least in the first two school years. This supports Hypothesis 3, though the evidence is weak.

4. Discussion

4.1. Summary and implications for educational research

Presently, there is a lack of empirical evidence on textbook effects because most research studies are based on cross-sectional designs, small sample sizes or the considered textbooks represented different curricula (frequently for the purpose of curriculum research). Our study contributes to the existing research by analyzing the effects of four different primary school textbooks representing the same curriculum on student achievement with a longitudinal design and a sound sample size. By re-analyzing an existing data set we could show that (1) mathematics teachers' textbook choice has a substantial effect on the students' arithmetic achievement in Grade 1 and 2 as well as their achievement in the content domains "number" and "pattern" in Grade 3 (between 11 and 24 percentage points explained variance), (2) there are substantial differences of the effects of the separate textbooks on students' achievement, and (3) there are indications that the effect of textbook choice is cumulative over the school years. These findings support the implications from the studies of Hadar (2017); Törnroos (2005) and Schmidt et al. (2001) that the mathematics textbook, as a medium for learning opportunities, not only has an effect on whether and how the mathematics content is taught, but also has an effect on student achievement.

Our findings differ from those of the cross-sectional study of van Steenbrugge et al. (2013), who did not find differential effects of textbooks on mathematics achievement. The contrary findings may be explained through the differences in the design. Although the study of Van Steenbrugge and colleagues had a similar sample size (1579 elementary school children), these children were distributed over six grades and used five different textbook series. If the textbook quality within one textbook series differed for different grades, or students from one grade have used a textbook from a different series in a previous grade, differential effects of the textbooks could have been reduced. The contradictory findings might also be explained by the fact that the textbooks in our sample strongly differ so that teachers address the same curricular topics in quite different ways. For example, the textbook "Einstern" in our sample suggests a specific learning trajectory and highly individualized learning activities. In comparison to other textbooks, the school classes taught by “Einstern” show a substantially lower achievement. Since textbooks have an effect on teaching practice (e.g., Krammer, 1985), and suggest a certain manner of use (e.g. Rabardel, 2002; Remillard, 2005; Rezat, 2008), students in our sample may have used “Einstern” mostly individually so that they did not get sufficient feedback to overcome their misconceptions. Like “Einstern” the textbook “Flex und Flo” is also structured in different booklets for each grade, which suggests a kind of macro structure of the textbook content. Since school classes taught by “Flex und Flo” also show a significantly lower student achievement in comparison to the textbook “Denken und Rechnen”, this might imply that such a booklet structure is less beneficial for the mathematics classroom than a traditional textbook structure.

The current study is based on a black box model and thus cannot provide evidence for an explanation for the identified effects of textbook choice. However, the effects on students' achievement are substantial and our study therefore supports the statement of Johansson (2016) that textbooks cannot merely be contemplated as an important tool for the teaching and learning of mathematics, but also as an obstacle.

A major implication for future research studies is to unravel the mechanisms causing the textbook effects. For this purpose, we suggest that future studies should examine the quality of the learning opportunities provided by the textbooks (cf. Hadar, 2017; Törnroos, 2005). Presently, we are not aware of instruments to assess overall textbook quality – even for well-structured school subjects like mathematics. The theory-driven development of valid quality indicators for subject-specific and subject-unspecific aspects of textbooks would be helpful to better determine the genuine influence of textbooks on student achievement. Furthermore, it would be worthwhile to examine if teachers' professional competences possibly moderate textbook effects on the instructional quality and on students' achievement. In addition, the way the textbook is implemented and possibly complemented by additional learning materials influences the quality of learning opportunities of the students and should therefore be explored in more detail. Apart from that one could assume that the effects found in this paper are also relevant for school subjects that are close to mathematics (STEM subjects), like science, technology and engineering. Therefore we recommend examining possible textbooks effects for those subjects. Furthermore it could be hypothesized that a close textbook orientation has an effect on student achievement. If this holds true effects of reading textbooks on student achievement can be expected in countries like Finland, Israel and the Netherlands (81% to 86% of the teachers use the textbook as a basis for instruction; Mullis, Martin, Foy, Arora et al., 2012; Mullis, Martin, Foy, Drucker et al., 2012) more likely than in Germany (62% of the teachers use the textbook as a basis for instruction; Mullis, Martin, Foy, Drucker et al., 2012).

As an unintended by-product, the results of our study suggest that the inclusion of textbooks as independent variables is worthwhile when analyzing effects of interventions. Our data set stemmed from the evaluation of two support programs for mathematics. Without controlling for the textbook used in the classroom, the effects of the support programs could not be identified as significant at the end of Grade 2 (Model 2 and 3 in Table 3).

Another interesting finding from this study is that teacher qualification had no significant influence on student achievement. This is in line with the previous finding from the German National Study 2016 at the end of Grade 4 (Rjosk et al., 2017). One possible reason for this is that the dichotomous variable is too distal to represent the teacher qualification adequately. Other measures like the professional competence of teachers are probably more revelatory control variables. Therefore, it is not surprising, that we found no difference in textbook usage between teachers with and without a formal qualification to teach mathematics, especially with both groups of teachers using the textbook extensively.

4.2. Educational implications

Our study provides empirical evidence that textbook choice has a substantial effect on student achievement even when textbooks are aligned with the same curriculum. These findings are highly relevant for educational practice.

The most important implications for educational practice concern the policy and the teacher training. For example, in Germany, the federal states can independently establish and disestablish a regulation for textbook approval so that some federal states have such regulations and others do not. In some other countries textbooks must be approved and licensed by the ministry of education (Section 1.1). The outcomes of this research suggest that textbook approval regulations based on theory-based quality indicators could be an instrument to avoid that textbooks with disadvantageous effects are used in schools and lead to an improvement of overall textbook quality. Furthermore the choice of high quality textbooks is a relatively easy, quick and inexpensive way to positively affect student achievement. Comparable effects through potential interventions on teacher or student level are related to higher expenses. The educational opportunities provided by textbooks in primary school might also affect student achievement in higher grades. The textbook effects on student achievement in the national test at the end of Grade 3 suggest, that students using certain books have advantages regarding attaining the National Educational Standards at the end of Grade 4. Future studies should address the long term effects of primary school textbooks as well as the role of secondary school textbooks on attaining the German National Educational Standards at the end of Grade 9 or international proficiency levels from studies like the Programme for International Student Assessment. If effects could be detected textbook choice could be an inexpensive way to improve the efficiency of the German educational system. Furthermore, teachers should be informed of possible textbook effects. Teacher education as well as teacher professional development can provide the opportunity to help pre- and in-service teachers reflecting on learning opportunities in textbooks and compensating an inadequate representation of teaching content.

4.3. Limitations

There are several limitations of our study. Since we re-analyzed an existing data set we were not able to administer specific instruments for our research. In particular, the questionnaires do not provide fine-grained data on the implementation of the teaching content or the teachers' professional knowledge. As a consequence this study could not unravel the black box on how the textbook influences student achievement and if teachers' professional knowledge possibly moderates the textbook effects. It could be worthwhile analyzing how the teachers implement the learning opportunities from the textbooks and how some teachers eventually compensate for inadequate learning opportunities in textbooks. Another limitation of this study is that we cannot exclude the possibility that teachers with certain characteristics (e.g., beliefs on teaching and learning) chose textbooks that match their teaching style. In the German federal state of Schleswig-Holstein, where the sample was drawn, the Education Act prescribes that the group of mathematics teachers of a school (“Fachkonferenz”) decides which mathematics textbook is used by all teachers in the respective school. Therefore, we do not expect that the textbooks in this study are chosen independently by the teachers of our sample and that teacher and textbook characteristics are strongly confounded. Finally, for the analysis of the achievement data for Grade 3 only sum scores of the national mathematics test for the topics numbers and patterns were available. This caused restrictions to the statistical data analyses because it was not possible to measure the longitudinal achievement development of the students on an unidimensional scale (like for the first two grades). Hence, there is a limitation for the analysis of the genuine effect of textbook choice in Grade 3 by controlling for the arithmetic achievement at the end of Grade 2 because of the different test content. This might have caused the non-significant textbook effect in Model 4 in Table 4. Furthermore, in Germany the national mathematics test in Grade 3 is administered and scored by the teachers. Therefore the data quality might be reduced.

4.4. Conclusion

Most obviously this study contributes to research of textbook effects on students' mathematics achievement. Because the analyzed textbooks mirror the same curriculum, the presented effects on student achievement are not influenced by different curricula but by different interpretations of a curriculum. The longitudinal data indicate the stability of textbook effects from the end of Grade 1 up to the end of Grade 3 and gives first hints for cumulative effects. The findings imply that in educational research and practice mathematics textbooks should be considered as a substantial factor for student learning.

Conflict of interest

None.

Acknowledgements

We would like to thank Dr. Inga Niedermeyer for her preparatory work that greatly assisted the research. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Enders, C. K. (2010). *Applied missing data analysis.* New York: Guilford Publications.
Fan, L., Zhu, Y., & Miao, Z. (2013). Textbook research in mathematics education: Development status and directions. *Zentralblatt für Didaktik der Mathematik, 45*(5), 633–646. https://doi.org/10.1007/s11858-013-0539-x.
Hadar, L. L. (2017). Opportunities to learn: Mathematics textbooks and students' achievements. *Studies in Educational Evaluation, 55*, 153–166.
Herbel-Eisenmann, B. A. (2007). From intended curriculum to written curriculum: Examining the "voice" of a mathematics textbook. *Journal for Research in Mathematics Education, 38*(4), 344–369.
Johansson, M. (2016). *Teaching mathematics with textbooks: A classroom and curricular perspective.* Retrieved fromLuleå: Luleå University of Technology. http://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-25802.
Jones, D. L., Brown, M., Dunkle, A., Hixon, L., Yorder, N., & Silbernick, Z. (2015). The statistical content of elementary school mathematics textbooks. *Journal of Statistics Education, 23*(3), 1–22.
Koedel, C., Morgan, D. L., Polikoff, S., Hardaway, T., & Wrabel, S. L. (2017). Mathematics curriculum effects on student achievement in California. *AERA Open, 3*(1), 1–22.
Kolovou, A., van den Heuvel-Panhuizen, M., & Bakker, A. (2009). Non-routine problem solving tasks in primary school mathematics textbooks: – A needle in a Haystack. *Mediterranean Journal for Research in Mathematics Education, 8*(2), 31–68.
Krammer, H. P. (1985). The textbook as classroom context variable. *Teaching and Teacher Education, 1*(4), 273–278. https://doi.org/10.1016/0742-051X(85)90015-0.
Lepik, M., Grevholm, B., & Viholainen, A. (2015). Using textbooks in The mathematics classroom – the teachers' view. *Nordic Studies in Mathematics Education, 20*(3–4), 129–156.
Lipowsky, F., Faust, G., & Karst, K. (Eds.). (2011). *Dokumentation der Erhebungsinstrumente des Projekts, Persönlichkeits- und Lernentwicklung von Grundschulkindern "(PERLE) - Teil 2. Materialien zur Bildungsforschung, Frankfurt am Main.* Germany: GFPF.
Lorenz, J. H. (2007). *HaRet – Hamburger Rechentest für Klasse 1.* Hamburg, Germany: Freie und Hansestadt Hamburg.
Mannhaupt, G. (2013). *MÜSC – Münsteraner Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten.* Berlin, Germany: Cornelsen.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.
Mayer, R. E., Sims, V., & Tajika, H. (1995). A comparison of how textbooks teach mathematical problem solving in Japan and the United States. *American Educational Research Journal, 32*(2), 443–460.
Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196.
Mullis, I. V., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1997). *mathematics achievement in the primary school years: IEA's Third International mathematics and science study (TIMSS).* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics.* Chestnut Hill, MA, Amsterdam: TIMMS International Association for the Evaluation of Educational Achievement (IEA).
Mullis, I. V., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *TIMSS 2011 international results in reading.* Chestnut Hill, MA, Amsterdam: TIMMS International Association for the Evaluation of Educational Achievement (IEA).
Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's Guide* (7th edn). Los Angeles, CA: Muthén and Muthén.
Pepin, B., & Haggarty, L. (2001). Mathematics textbooks and their use in English, French and German classrooms: A way to understand teaching and learning cultures. *Zentralblatt für Didaktik der Mathematik, 33*(5), 158–175.
Rabardel, P. (2002). *People and technology: A cognitive approach to contemporary instruments.* Retrieved 17.01.2018 fromParis: Université Paris. https://hal.archives-ouvertes.fr/hal-01020705.
Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research, 75*(2), 211–246. https://doi.org/10.3102/00346543075002211.
Rezat, S. (2008). Die Struktur von Mathematikschulbüchern [The structure of mathematics textbooks]. *Journal für Mathematik-Didaktik, 29*(1), 46–67. https://doi.org/10.1007/BF03339361.
Rjosk, C., Hoffmann, L., Richter, D., Marx, A., & Gresch, C. (2017). Qualifikation von Lehrkräften und Einschätzungen zum gemeinsamen Unterricht von Kindern mit und Kindern ohne sonderpädagogischen Förderbedarf [Teacher qualification and evaluation of joint instruction of children with and without special needs]. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich, & N. Haag (Eds.). *IQB-Bildungstrend 2016: Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe in zweiten Ländervergleich* (pp. 335–353). Münster: Waxmann.
Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H., Wiley, D. E., Cogan, L. S., et al. (2001). *Why schools matter: A cross-national comparison of curriculum and learning. The Jossey-Bass education series.* San Francisco, CA: Jossey-Bass.
Schmidt, W. H., Raizen, S. A., Britton, E. D., Bianchi, L. J., & Wolfe, R. G. (1997). *A Cross-national investigation of curricular intentions in school mathematics.* Dordrecht: Kluwer Academic.
Stein, M., Remillard, J., & Smith, M. (2007). How curriculum influences students' learning. In F. K. Lester (Ed.). *Second handbook of research on mathematics teaching and learning: A project of the national council of teachers of mathematics* (pp. 319–370). Charlotte, N.C: Information Age Pub.
Tarr, J. E., Reys, R. E., Reys, B. J., Chávez, Ó., Shih, J., & Osterlind, S. J. (2008). The impact of middle-grades mathematics curricula and the classroom learning environment on student achievement. *Journal for Research in Mathematics Education, 39*(3), 247–280.
Törnroos, J. (2005). Mathematics textbooks, opportunity to learn and student achievement. *Studies in Educational Evaluation, 31*(4), 315–327. https://doi.org/10.1016/j.stueduc.2005.11.005.
Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen, & K. Elliot (Eds.). *But what does it mean? The use of effect sizes in educational research* (pp. 55–66). London: National Foundation of Educational Research.
Valverde, G., Bianchi, L. J., Wolfe, R., Schmidt, W. H., & Houang, R. T. (2002). *According to the book: Using TIMSS to investigate the translation of policy into practice through the world of textbooks.* Netherlands: Springer.
van Steenbrugge, H., Valcke, M., & Desoete, A. (2013). Teachers' views of mathematics textbook series in Flanders: Does it (not) matter which mathematics textbook series schools choose? *Journal of Curriculum Studies, 45*(3), 322–353. https://doi.org/10.1080/00220272.2012.713995.
von Davier, M., Gonzalez, E., & Mislevy, R. (2009). Plausible values: What are they and why do we need them? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2*, 9–36.
Weiß, R. H., & Osterland, J. (2013). *CFT 1-R – Grundintelligenztest Skala 1 Revision.* Göttingen, Germany: Hogrefe.
Zimmer-Müller, M. (2013). *Zehn Jahre Vergleichsarbeiten: Eine Zwischenbilanz aus verschiedenen Perspektiven [Ten Years of the national mathematics test: A mid-term evaluation from different perspectives]. Empirische Pädagogik, Vol. 27,* Landau: Verlag Empirische Pädagogik4.