

Empirische Sonderpädagogik, 2022, Nr. 3, S. 211-227
ISSN 1869-4845 (Print) · ISSN 1869-4934 (ebook)

Learning from null results: An analysis of the unexpected findings of a mathematical intervention study conducted in inclusive classrooms

Urs Grob, Elisabeth Moser Opitz & Meret Stöckli

Universität Zürich

Abstract

This paper presents an additional, more detailed, analysis of the null results and unexpected outcome of a published intervention study. Pfister, Stöckli et al. (2015) report the results of an intervention study in mathematics in inclusive classrooms. In that study the general education teacher implemented an intervention aimed at supporting low achievers in mathematics lessons in 58 inclusive classrooms with 888 third grade students. Two experimental groups ($n = 37$ teachers) worked with the program, the third group (group^{CONTROL}; $n = 21$ teachers) continued to teach 'as usual'. The experimental group^{MAT} ($n = 16$ teachers) was given the program material and the experimental group^{MEET} ($n = 21$ teachers) received the material and had two in-service training sessions. Contrary to theory-based predictions, group^{MAT} outperformed group^{MEET}. To try to understand this finding, in the present study we investigated the effects of the three treatment conditions on students with different levels of math achievement at t1 by using the data to estimate random slope models with cross-level interactions. Contrary to our expectations, high achieving students in group^{MAT} had significantly greater learning gains than those in group^{MEET}. Control variables at the individual and class level could not explain this outcome. The analysis highlighted the methodological challenges of studies conducted in complex inclusive education settings and raised the question of whether targeted support in an inclusive setting, without individual support outside of the classroom, can meet the needs of low achieving students in mathematics.

Keywords: intervention study, null effects, low achievers in mathematics, inclusive education

Lernen von Null-Effekten: Eine Analyse von unerwarteten Effekten einer Interventionsstudie in inklusiven Klassen

Zusammenfassung

In der Publikation von Pfister, Stöckli et al. (2015) wurden unerwartete Ergebnisse und fehlende Effekte einer unterrichtsintegrierten Intervention für rechenschwache Schüler:innen in inklusiven Klassen berichtet. Im vorliegenden Artikel werden zusätzliche und detailliertere Analysen präsentiert, die dazu beitragen, die Ergebnisse zu verstehen und einzuordnen. In der Studie mit ($N = 58$ inklusive Klassen, $N = 888$ Drittklässler:innen) wurde eine

unterrichtsintegrierte Intervention zur Förderung von rechenschwachen Schüler:innen von den Regellehrkräften durchgeführt. Zwei Experimentalgruppen ($n = 37$ Lehrkräfte) erhielten die Fördermaterialien, die Lehrkräfte ($n = 21$) in der Kontrollgruppe (Gruppe^{CONTROL}) führte ihren normalen Mathematikunterricht durch. Die Experimentalgruppe Gruppe^{MAT} ($n = 16$ Lehrkräfte) erhielt die Fördermaterialien, die Experimentalgruppe Gruppe^{MEET} ($n = 21$ Lehrkräfte) bekam zusätzlich zu den Materialien ein Fortbildungsangebot (zwei Nachmittage). Entgegen den Erwartungen machten die Lernenden in der Gruppe^{MAT} größere Leistungsfortschritte als die Gruppe^{MEET}. Um dieses Ergebnis zu verstehen, wurde in der hier präsentierten Studie der Einfluss der unterschiedlichen Interventionsformen auf Schüler:innen mit unterschiedlichen mathematischen Vorkenntnissen untersucht. Berechnet wurden Random Slope Modelle mit Cross-Level Interaktionen. Entgegen den Erwartungen machten Lernende mit hohen Vorkenntnissen in der Gruppe^{MAT} signifikant größere Leistungsfortschritte als vergleichbare Lernende in der Gruppe^{MEET}. Die Kontrollvariablen auf Individual- und Klassenebene konnten dieses Ergebnis nicht erklären. Dieses Resultat weist erstens darauf hin, dass es sehr herausfordernd ist, inklusiven Unterricht zu untersuchen und die verschiedenen Einflussvariablen zu kontrollieren. Zudem stellt sich die Frage, ob eine rein unterrichtsintegrierte Förderung den Bedürfnissen von rechenschwachen Lernenden gerecht wird und ob diese nicht auch eine individualisiertere Förderung außerhalb des Regelunterrichts benötigen würden.

Schlagwörter: Interventionsstudie, Null-Effekte, rechenschwache Schüler:innen, inklusiver Unterricht

There has been much research into how best to support students with below average achievement in mathematics. Researchers have shown that these students need extra support to develop a good grasp of basic mathematical knowledge and key concepts (e.g. Andersson, 2010; Chan et al., 2017) and that organized programs with explicit instruction are generally the best way to achieve this (e.g. Chodura et al., 2015; Gersten et al., 2009; Ise et al., 2012; Stevens et al., 2018). Often, intervention programs for low achievers in mathematics involve small group settings or individual support and meta-analyses have confirmed that individual, one-to-one, support is beneficial (Chodura et al., 2015; Ise et al., 2012). The results of remedial interventions in full classroom settings have been mixed (e.g. Fuchs et al., 2015; Jitendra et al., 2017; Moser Opitz et al., 2017; Zhang & Xin, 2012) and success seems to depend on other factors such as the support of a special education teacher

(Bottge et al., 2017). But while one-to-one support and small group interventions may be more effective than support in an inclusive setting, they often take place outside of the regular classroom which means they work against the general objectives of preventing exclusion and supporting inclusivity in education (e.g. Scherer et al., 2016; Feuser, 1989) and have disadvantages for student social participation (Wiener & Tar-diff, 2004).

Providing individual support in the highly heterogeneous environment of an inclusive classroom is challenging for teachers. Making an empirical evaluation of any intervention is also complex because many factors can influence the outcome of any inclusive remedial program: Student characteristics (e.g. prior knowledge, intelligence, socio-economic background, language proficiency; e.g. Powell et al., 2017; Wood et al., 2020), implementation quality (e.g. Hagermoser Sanetti & Fallon, 2011), aspects

of teaching quality such as classroom management or student support (e.g. Praetorius et al., 2018), and class composition (e.g. Zurbriggen, 2016). Because it is difficult to control for all of these complicating factors, there is a strong likelihood that an intervention study in an inclusive setting will have null or unexpected effects. Studies with null results – and those with unexpected or unwanted results – are not often published (Conlin et al., 2019) and Gage et al. (2017) found that less than 50% of the meta-analyses in special education discuss this publication bias. But as Jacob et al. (2019, p. 581) points out, null results “have the potential to yield valuable information beyond simply ‘this didn’t work.’”

Null-effects and an inverted difference between the two experimental groups were the finding of an analysis of the outcomes of a remedial program in mathematics for low achievers reported in Pfister, Stöckli, et al. (2015) and Stöckli (2019). A multi-level regression analysis conducted by Stöckli (2019) also showed that the intervention had no significant impact on low achievers in mathematics, the intervention’s target group. In this paper we investigate the data provided in Pfister, Stöckli, et al. (2015) and Stöckli (2019) to see if additional analyses looking at the potential differential effects of the treatment conditions of the program on students with specific mathematical achievement profiles (low achievers, high achievers) can help to elucidate the reasons for the unexpected outcome.

Effective strategies for supporting low achievers in mathematics

Students who struggle to learn mathematics are variously referred as students with learning difficulties in mathematics, students with mathematical learning disabilities, or low achievers in mathematics. These students, who we collectively term *low achievers in mathematics*, are characterized by having large gaps in their mathematical knowledge

relative to their peers. They especially have difficulty understanding the concept of the base-10 number system (see e.g. Andersson, 2010; Chan et al., 2017) and solving word problems (e.g. Zhang & Xin, 2012). On a procedural level, they have problems with counting by groups (e.g. Desoete et al., 2009; Moser Opitz, 2013) and retrieving facts (e.g. Andersson, 2010). Therefore, many intervention programs address these issues.

Empirical evidence from meta-analyses shows that structured, organized, programs and explicit instruction improve the performance of low achievers in mathematics (e.g. Chodura et al., 2015; Gersten et al., 2009; Stevens et al., 2018). The studies included in the meta-analyses are, however, very heterogeneous. The numbers of participants, characteristics and content of the intervention, and standardized measures used all vary and there is also a high effect size variance (Chodura et al., 2015; Stevens et al. 2018). The results of studies exploring which setting may be most suitable for supporting low achievers in mathematics are also inconclusive. Ise et al. (2012) found that one-to-one training was more beneficial than small group support and Chodura et al. (2015) reported promising results for one-to-one interventions, but only for children with at-risk dyscalculia. However, Jitendra et al. (2021) found a positive effect for an intervention carried out with groups of two or three children at a time and Stevens et al. (2018) found that group size had no impact.

Another point to consider is that one-to-one support and small group interventions are often provided in separate resource rooms and it is not clear if effective targeted interventions are possible in inclusive settings. Some studies show that supporting students in a separate setting can be advantageous. In a study by Bottge et al. (2017), students receiving an intervention on fractions in a resource room had higher achievement gains than those who were taught in the inclusive setting. Fuchs et al. (2015) found that students receiving spe-

cialized education in small groups significantly outperformed students in a co-taught inclusive classroom. Other studies indicate that interventions in inclusive settings could be more effective. A meta-analysis focused on word problem solving skills found smaller effect sizes for interventions in special educational settings compared to those in inclusive settings (Zang & Xin, 2012). In a study focused on the children's understanding of basic arithmetical concepts, Moser Opitz et al. (2017) reported that an intervention program to improve the mathematical achievement gain of low achievers in fifth grade who were partially integrated in the regular classroom had significant positive effects.

Interventions which involve taking students outside of the regular classroom for extra tutoring also have their critics. According to Hinz and Köpfer (2016), Feuser (1989), and Scherer et al. (2016) such practices are against the principles of inclusive education, which aims to offer joint learning situations for all students by having a high level of differentiation within the classroom. Removing students with special educational needs from the classroom for support lessons can result in harmful labels and hinder their social participation and development. Wiener and Tardiff (2004) showed that regular support outside the classroom has a negative effect on the social acceptance of students.

Hoping to address some of these issues, the intervention study asked general education teachers to implement a modified version of a program developed by Freesemann (2014) in inclusive third grade mathematics classes (see Pfister, Stöckli, et al. 2015; Pfister, Moser Opitz, & Pauli, 2015; Stöckli, 2019 for details). Two experimental groups (group^{MEET}: program and in-service training and group^{MAT}: program only) executed the program while the control group continued as usual. Based on the results of a study by Moser Opitz et al. (2017), it was hypothesized that the low achievers in mathematics in the experimental classrooms would

make greater gains than those in the control group. Because structured in-service training has been shown to be beneficial (Lipowsky, 2004), low achieving students in group^{MEET} were expected to make more progress than those in group^{MAT}.

Results reported by Pfister, Stöckli, et al. (2015) and Stöckli (2019) revealed null and unexpected effects. When controlling for math achievement t1, IQ, SES, age, and gender at level 1 a multilevel analysis found no intervention effect for group^{MEET} and group^{MAT} at t2 (Pfister, Stöckli, et al., 2015). Contrary to expectations, group^{MAT} outperformed group^{MEET}. When selected control variables were added at level 2, there was only a small intervention effect for group^{MAT} compared to the control group (Stöckli, 2019). This suggested that the outcome of the study might have been dependent on the achievement level of the students at t1.

Research questions

The results presented in Pfister, Stöckli, et al. (2015) and Stöckli (2019) indicate that the intervention might have only been suitable and beneficial for students with specific mathematical achievement profiles. Therefore, this new analysis explores whether the intervention interacted with the mathematical achievement level of the students in each class at t1 to affect the study outcomes, leading to the following research questions:

RQ1) Do the treatment conditions of the intervention program have a differential effect on students with a low level of mathematical achievement at t1 in each class?

RQ2) Do the treatment conditions of the intervention program have a differential effect on students with a high level of mathematical achievement at t1 in each class?

Method

Design

A remedial program to support low achievers in mathematics was implemented by the general education teacher during the daily mathematics lesson of 888 third graders in 58 inclusive classrooms over a period of 20 weeks. The program was based on an intervention program by Freesemann (2014) which has been shown to significantly improve the learning gain of low achievers in mathematics in Grade 5 (Moser Opitz et al., 2017). It included the following components (for a detailed description and examples see Pfister, Moser Opitz, & Pauli, 2015; Stöckli et al., 2014):

- 30 highly structured lesson plans on the base-10 number system (grouping, de-grouping, place value), number line, and flexible addition/subtraction strategies). The plans showed the teachers how to adapt the Grade 3 curriculum for low achievers in mathematics and included suggestions for how to replace textbook pages and select suitable content and tasks for these students. The plans had suggestions for how to review existing knowledge (e.g. flexible addition strategies up to 100) and add new knowledge (e.g. flexible addition strategies up to 1000).
- Each lesson plan consisted of three phases: Introduction to the whole class, individualized work phase on “core concepts”, and classroom discussion. Suggestions for how to differentiate the learning material for students with different achievement levels in each phase were part of the lesson plans
- Flashcards with different achievement levels for work phase or training sessions: Counting forwards and backwards by steps, and mental calculation.
- The intervention followed the concept of scaffolding and included pre-formulated questions and hints for the teachers.

All classes followed the Grade 3 curriculum in arithmetic during the intervention phase (numbers to 1000, flexible addition and subtraction strategies) using one of two very similar textbooks. The program was implemented under three conditions and the teachers were randomly assigned to the three groups: Two experimental groups ($n = 37$ teachers) worked with the program, the third group (group^{CONTROL}; $n = 21$ teachers) continued teaching mathematics lessons as usual following the textbook. The teachers of the experimental groups attended a three-hour meeting at the beginning of the school year where they were introduced to the program and provided with lesson plans, worksheets, flashcards, and manipulatives. The teachers in group^{MAT} ($n = 16$) only attended this meeting. The teachers in the second experimental group (group^{MEET}, $n = 21$) had two additional training sessions during which they were given examples of best-practice and were able to discuss their experiences of using the program.

Mathematical achievement was tested at the beginning of the school year (pre-test, t1) and after the intervention had ended, 6 months later (post-test, t2)¹. Information on socio-economic status (SES) was also collected twice, at t1 and t2. Data on IQ, first language, German language comprehension and special educational needs in mathematics (SEN^{MATH}) were collected at t1. To monitor implementation fidelity, the teachers were asked to submit logs showing how they implemented the lesson plans and one mathematics lesson in each of the classes in the experimental groups was video recorded. Teacher behavior was analyzed using a high-inference rating (Pfister, Moser Opitz, & Pauli, 2015). The video recordings revealed that in 12 classes (group^{MEET}, group^{MAT}) a second teacher was sometimes present in the mathematics lessons. However, it is not known if this was a trained special education teacher (SET) and how he/she was involved in the mathematics teaching.

¹ Mathematical achievement was also tested at the end of grade 3 and one year later, at the end of grade 4. To reduce complexity, and because the null effect remained stable, we only report the results for t1 and t2.

Measures

Mathematical achievement was assessed using two standardized math tests which were in development for publication at the time of measurement. The tests measure whether students have acquired the basic arithmetic knowledge expected at their age/grade. The pre-test (t1) was conducted using 30 items from BASIS-MATH-G 2⁺ (WLE-reliability .84; Moser Opitz, Stöckli, et al., 2020). The post-test (t2) was conducted using 41 items from BASIS-MATH-G 3⁺ (WLE reliability .89; Moser Opitz et al., 2019).

IQ was assessed using CFT 1 (Cattell et al., 1997) and the SES book-task (Paulus, 2009). Data on language comprehension in German and SEN^{MATH} were collected by asking teachers to complete a questionnaire at t1.

Sample

The sample comprised 888 third graders (50% female) in 58 classes² in nine German-speaking cantons of Switzerland. The average cluster size was 15.3 students per class ($SD = 5.2$) and the mean age of the participating children at t1 was 8.7 years ($SD = 0.5$). ICC (1) of the math test score was 9.6% at t1 and 13.5% at t2. Missing values proportions ranged from 0.0% (student's gender) to 2.0% (math score at t2).

Data analysis

Earlier analyses by members of the project team applied multilevel analysis using HLM 7 (Raudenbush et al., 2013), that focused on the fixed effects of the treatment conditions at class level, controlling for the linear effects of the students' mathematical achievement at t1 at the individual level (Pfister, Stöckli, et al., 2015) and class level (Stöckli, 2019). In order to identify possible differential effects of the three treatment conditions (group^{MEET}, group^{MAT}, group^{CONTROL}) on stu-

dents with different levels of math achievement at t1, two versions of a random slope model with cross-level-interactions were estimated in Mplus 8.7 (Muthén & Muthén, 2017).

In a first step, as a baseline, a two-level random intercept model with math achievement t2 as dependent variable at both levels was run (model 1). At level 1, math achievement t1, IQ, SES, age, gender, SEN^{MATH}, and German comprehension were included as predictors. At level 2, the model included mean math achievement t1 and two level-2 dummy variables representing the treatment conditions as predictors. Math achievement scores at t1 and t2 were decomposed into latent level-specific variance components (Lüdtke et al., 2008). To assess all contrasts between the three treatment groups, this model was estimated in two versions based on different reference groups: In model 1a the reference group was group^{CONTROL}, in model 1b the reference group was group^{MEET}.

To answer the first research question the baseline model, 1 a/b, was augmented by a random slope for the additional effect arising from a student being part of the *lowest* quartile in math achievement *in their class* at t1 while controlling for the linear effect of math achievement at t1. At level 2, this random slope was regressed on the latent mean math achievement at t1 and two level-2 dummy variables, representing the treatment conditions. Again, different reference groups were defined for models 2a and 2b.

Finally, in order to answer the second research question the random slope part of model 2 a/b was modified. In the resulting model 3 a/b the random slope was generated by regressing math achievement at t2 on a binary variable which represented being part of the *highest* math achievement quartile *in their class* at t1.

Therefore, models 2 a/b and 3 a/b test whether there are differential treatment effects that depend on the relative achievement-related position of students *in their*

² There were 61 classes in the beginning. One teacher decided to leave the project and two classes had to be excluded because the quality of the intervention did not meet the study standards.

class; whether weak or strong learners benefit disproportionately from each treatment condition.

The notation of the three models is as follows:

Model 1a³:

$$\begin{aligned} \text{math_achievement_t2}_{ic} = & \beta_{0c} + \beta_1(\text{math_achievement_t1}_{ic} - \text{mean_math_achievement_t1}_c) \\ & + \beta_2IQ_i + \beta_3SES_i + \beta_4age_i + \beta_5gender_i + \beta_6\text{math_special_needs}_i \\ & + \beta_7\text{comprehension_of_German}_i + \varepsilon_{ic} \end{aligned}$$

$$\begin{aligned} \beta_{0c} = & Y_{00} + Y_{01}\text{mean_math_achievement_t1}_c + Y_{02}\text{group_meet}_c + Y_{03}\text{group_mat}_c + u_{0c} \end{aligned}$$

In a slight deviation from this classical notation, math achievement at t1 and t2 were decomposed into latent within-class and between-class components in all models. At the individual level, this essentially corresponds to centering the values around each group mean. Therefore, the manifest difference term $\text{math_achievement_t1}_{ic} - \text{mean_math_achievement_t1}_c$ is used in the notation above.

Model 2a:

$$\begin{aligned} \text{math_achievement_t2}_{ic} = & \beta_{0c} + \beta_1(\text{math_achievement_t1}_{ic} - \text{mean_math_achievement_t1}_c) \\ & + \beta_{2c}\text{math_achievement_t1_Quartile1}_{ic} + \beta_3IQ_i + \beta_4SES_i + \beta_5age_i + \beta_6gender_i + \beta_7\text{math_special_needs}_i \\ & + \beta_8\text{comprehension_of_German}_i + \varepsilon_{ic} \end{aligned}$$

$$\begin{aligned} \beta_{0c} = & Y_{00} + Y_{01}\text{mean_math_achievement_t1}_c + Y_{02}\text{group_meet}_c + Y_{03}\text{group_mat}_c + u_{0c} \end{aligned}$$

$$\begin{aligned} \beta_{2c} = & Y_{20} + Y_{21}\text{mean_math_achievement_t1}_c + Y_{22}\text{group_meet}_c + Y_{23}\text{group_mat}_c + u_{2c} \end{aligned}$$

In models 2a and 2b, the random slope relates to a binary level 1 variable ($\text{math_achievement_t1_Quartile1}$) indicating that the students belong to the first (lowest) math achievement quartile in each class at t1 (value: 1) vs. being part of one of the other three quartiles (value: 0). The percentage of students in the first quartile does not vary significantly between classes ($M = 0.249$, $SD = 0.044$, $F(57, 818) = 0.108$, $p = 1.000$).

Model 3a:

$$\begin{aligned} \text{math_achievement_t2}_{ic} = & \beta_{0c} + \beta_1(\text{math_achievement_t1}_{ic} - \text{mean_math_achievement_t1}_c) \\ & + \beta_{2c}\text{math_achievement_t1_Quartile4}_{ic} + \beta_3IQ_i + \beta_4SES_i + \beta_5age_i + \beta_6gender_i + \beta_7\text{math_special_needs}_i \\ & + \beta_8\text{comprehension_of_German}_i + \varepsilon_{ic} \end{aligned}$$

$$\begin{aligned} \beta_{0c} = & Y_{00} + Y_{01}\text{mean_math_achievement_t1}_c + Y_{02}\text{group_meet}_c + Y_{03}\text{group_mat}_c + u_{0c} \end{aligned}$$

$$\begin{aligned} \beta_{2c} = & Y_{20} + Y_{21}\text{mean_math_achievement_t1}_c + Y_{22}\text{group_meet}_c + Y_{23}\text{group_mat}_c + u_{2c} \end{aligned}$$

Models 3a and 3b are almost identical to models 2a and 2b. The only difference is that the binary indicator of the achievement related position in each class ($\text{math_achievement_t1_Quartile4}$) reflects being part of the fourth (highest) math achievement quartile in each class at t1. The percentage of students in the fourth quartile does not vary significantly between classes ($M = 0.251$, $SD = 0.049$, $F(57, 818) = 0.160$, $p = 1.000$).

In models 2 a/b and 3 a/b, the random slope was allowed to correlate with the residual of math achievement at t2. In all models, mean values of the IQ, SES, age, gender, SEN^{MATH} and German comprehension were not included at level 2. Given the limited number of cases at level 2, the inclusion of more level 2 predictors would have led to an increased risk of overfitting. Due to slight aberrations from a normal distribution, Maximum Likelihood Robust Estimator (MLR) was chosen as the estimator for all models. As a result, there were no standardized effects and explained variances.

Results

The results of models 1a and 1b are listed in Table 1. Because model 1 includes only fixed effects and random intercepts and has no random slope, it acts as a baseline.

Models 1a and 1b show – with identical coefficients for all level-1 predictors (cf. the within part of table 1) – that at the individual level, math achievement at t1 predicted math achievement at t2 with a high de-

3 In models 1b, 2b, and 3b, the reference group was $\text{group}^{\text{MEET}}$ instead of $\text{group}^{\text{CONTROL}}$, which allowed us to assess the contrast between $\text{group}^{\text{MEET}}$ and $\text{group}^{\text{MAT}}$.

Table 1
 Multilevel regression model for the prediction of t1 to t2 math achievement gain depending on the treatment condition (Models 1a & 1b)

	Model 1a			Model 1b				
	b	SE	t	p (2-sided)	b	SE	t	p (2-sided)
Within								
Math achievement t1	0.583	0.035	16.772	.000	0.583	0.035	16.772	.000
IQ	0.015	0.002	5.914	.000	0.015	0.002	5.914	.000
SES	-0.001	0.028	-0.027	.978	-0.001	0.028	-0.027	.978
Age (months)	0.002	0.005	0.497	.619	0.002	0.005	0.497	.619
Gender (f = 0, m = 1)	0.182	0.064	2.852	.004	0.182	0.064	2.852	.004
SEN ^{MATH} (no = 0, yes = 1)	-0.641	0.093	-6.916	.000	-0.641	0.093	-6.916	.000
German comprehension (good = 0, limited = 1)	-0.144	0.067	-2.139	.032	-0.144	0.067	-2.139	.032
Residual variance	0.710	0.036	19.605	.000	0.710	0.036	19.605	.000
Between (model 1a; ref = group^{CONTROL})								
Intercept math achievement t2	-1.287	0.641	-2.008	.045				
Math achievement t1	0.909	0.218	4.165	.000				
Group ^{MAT}	0.247	0.131	1.880	.060				
Group ^{MEET}	-0.137	0.110	-1.242	.214				
Residual variance	0.084	0.026	3.236	.001				
Between (model 1b; ref = group^{MEET})								
Intercept math achievement t2					-1.424	0.636	-2.240	.025
Mean math achievement t1					0.909	0.218	4.166	.000
Group ^{MAT}					0.384	0.138	2.783	.005
Group ^{CONTROL}					0.137	0.110	1.242	.214
Residual variance					0.084	0.026	3.236	.001

Notes: Non-standardized regression coefficients with standard errors based on MLR estimation
 Latent level-specific decomposition of mathematics score t1 and t2

gree of significance ($b = 0.583$, $t = 16.772$, $p = .000$). A one WLE unit (logit) increase at t1 resulted in an expected value at t2 that was, on average, 0.58 units higher.

Controlling for this stability effect, IQ ($b = 0.015$, $t = 5.914$, $p = .000$), gender ($b = 0.180$, $t = 2.852$, $p = .004$), SEN^{MATH} ($b = -0.642$, $t = -6.916$, $p = .000$), and German comprehension ($b = -0.144$, $t = -2.139$, $p = .032$) showed significant unique effects on the *change* in math achievement between t1 and t2. Students with a higher IQ, male students, students without SEN^{MATH} and without German language limitations had higher mathematical achievement gain from t1 to t2. No significant effects of level-1 predictors SES and age were found.

In the between part of both model 1a and model 1b, that is, at class level in each model, math achievement at t1 predicted math achievement at t2 with a high degree of significance ($b = 0.909$, $t = 4.165$, $p = .000$). A (latent) mean class value 1 WLE unit (logit) higher at t1 resulted in an expected mean class value that was, on average, 0.91 units higher at t2.

Third, testing for differential effects between the three treatment conditions, again at class level, resulted in only one significant contrast (cf. model 1b): Students in the classes in $group^{MAT}$ had a higher mathematical achievement gain than students in the classes in $group^{MEET}$ ($b = 0.384$, $t = 2.783$, $p = .005$). The unexpected greater improvement shown by students in $group^{MAT}$ compared to those in $group^{MEET}$, which was also highlighted by previous analyses (Pfister, Stöckli, et al., 2015; Stöckli, 2019), requires an explanation. To determine whether the intervention effect depended on the fit of the program with the needs of students with different achievement levels, models 2 and 3 test whether there is an interaction between the treatment condition and the individual achievement level relative to their class mean.

The results of models 2a and 2b, addressing RQ1, are listed in table 2. The models include a random slope for being part of

the lowest t1 math achievement quartile of each class and a cross-level-interaction with mean mathematics achievement level at t1 and treatment conditions.

In models 2a and 2b all coefficients at level 1 are identical and they are virtually identical with those in models 1a and 1b. Therefore, the focus is on the between part. The models show comparable fixed effects at class level for mean math achievement at t1. Adding the random effect makes the difference between the mean gain in math achievement of all classes in $group^{MAT}$ compared to those in $group^{CONTROL}$ significant ($b = 0.291$, $t = 2.224$, $p = .026$). In model 1, this effect was not significant, although the significance threshold was just narrowly missed ($b = 0.247$, $t = 1.880$, $p = .060$). Given the small difference in the error probability, this result has little relevance. The achievement gain related difference between the classes in $group^{MAT}$ and in $group^{MEET}$ (cf. model 2b) is significant, and the error probability ($b = 0.430$, $t = 3.000$, $p = .003$) is roughly the same as in model 1b ($b = 0.384$, $t = 2.783$, $p = .005$).

In the random part, no significant effect was found, in either model 2a or model 2b. This means that the difference in the t1 to t2 mathematical achievement gain between the students in the quartile with the lowest achievement in each class is not systematically different from the achievement gain of the other students in each class. It is slightly below zero, but not statistically significant. The between class variation of the t1 to t2 achievement gain of the students in the quartile with the lowest level of mathematical achievement in each class, compared to the three higher quartiles, also does not depend on the mean math achievement level in each class. Moreover, the relative achievement gain of the weakest learners in each class is not dependent on the type of treatment.

To address RQ2, models 3a and 3b included a random slope and a cross-level-interaction for students being part of the quartile with *highest achievement* in their class.

Table 2
 Multilevel regression model for the prediction of t1 to t2 math achievement gain depending on the treatment condition with cross-level-interaction for being part of the lowest t1 math achievement quartile in a class (Models 2a & 2b)

	Model 2a			Model 2b				
	<i>b</i>	SE	<i>t</i>	<i>p</i> (2-sided)	<i>b</i>	SE	<i>t</i>	<i>p</i> (2-sided)
Within								
Math achievement t1	0.584	0.035	16.711	.000	0.584	0.035	16.711	.000
IQ	0.014	0.002	5.861	.000	0.014	0.002	5.861	.000
SES	-0.003	0.028	-0.111	.911	-0.003	0.028	-0.111	.911
Age (months)	0.003	0.005	0.478	.633	0.003	0.005	0.478	.633
Gender (f = 0, m = 1)	0.179	0.064	2.798	.005	0.179	0.064	2.798	.005
SEN ^{MATH} (no = 0, yes = 1)	-0.637	0.091	-7.000	.000	-0.637	0.091	-7.000	.000
German comprehension (good = 0, limited = 1)	-0.132	0.066	-2.010	.044	-0.132	0.066	-2.010	.044
Residual variance	0.703	0.038	18.441	.000	0.703	0.038	18.441	.000
Between (model 1a; ref = group^{CONTROL})								
Intercept math achievement t2	-1.216	0.680	-1.787	.074				
Math achievement t1	0.786	0.182	4.308	.000				
Group ^{MAT}	0.291	0.131	2.224	.026				
Group ^{WAI}	-0.139	0.108	-1.288	.198				
Residual variance intercept	0.076	0.024	3.150	.002				
Cross-level interaction: effect on rand. slope (Math ach. t2 regressed on math ach. Q1 t1)								
Intercept random slope	-0.140	0.202	-0.693	.489				
Math achievement t1	0.328	0.235	1.395	.163				
Group ^{MAT}	-0.114	0.181	-0.633	.527				
Group ^{WAI}	0.058	0.187	0.310	.757				
Residual variance random slope	0.013	0.054	0.245	.806				
Between (model 1b; ref = group^{WAI})								
Intercept math achievement t2					-1.355	0.662	-2.047	.041
Math achievement t1					0.786	0.182	4.308	.000
Group ^{MAT}					0.430	0.142	3.000	.003
Group ^{CONTROL}					0.139	0.108	1.288	.198
Residual variance intercept					0.076	0.024	3.150	.002
Cross-level interaction: effects on rand. slope (Math ach. t2 regressed on math ach. Q1 t1)								
Intercept random slope					-0.082	0.193	-0.425	.671
Math achievement t1					0.328	0.235	1.394	.163
Group ^{MAT}					-0.173	0.176	-0.983	.326
Group ^{CONTROL}					-0.058	0.187	-0.310	.757
Residual variance random slope					0.013	0.054	0.245	.806

Notes: Non-standardized regression coefficients with standard errors based on MLR estimation
 Latent level-specific decomposition of mathematics score t1 and t2

Table 3 Multilevel regression model for the prediction of t1 to t2 gain in math achievement depending on the treatment condition with cross-level-interaction for being part of the highest t1 math achievement quartile per class (Models 3a & 3b)

	Model 3a			Model 3b				
	<i>b</i>	SE	<i>t</i>	<i>p</i> (2-sided)	<i>b</i>	SE	<i>t</i>	<i>p</i> (2-sided)
Within								
Math achievement t1	0.615	0.051	12.040	0.000	0.615	0.051	12.039	0.000
IQ	0.012	0.003	4.860	0.000	0.012	0.003	4.860	0.000
SES	-0.011	0.029	-0.389	0.697	-0.011	0.029	-0.389	0.697
Age (months)	-0.006	0.005	-1.080	0.280	-0.006	0.005	-1.079	0.281
Gender (f = 0, m = 1)	0.180	0.064	2.798	0.005	0.180	0.064	2.798	0.005
SEN ^{MATH} (no = 0, yes = 1)	-0.632	0.092	-6.850	0.000	-0.632	0.092	-6.850	0.000
German comprehension (good = 0, limited = 1)	-0.125	0.067	-1.858	0.063	-0.125	0.067	-1.858	0.063
Residual variance	0.704	0.037	19.084	0.000	0.704	0.037	19.084	0.000
Between (model 1a; ref = group^{CONTROL})								
Intercept math achievement t2	-0.136	0.700	-0.194	0.846				
Math achievement t1	0.928	0.167	5.565	0.000				
Group ^{MAT}	0.231	0.136	1.700	0.089				
Group ^{MEET}	-0.083	0.120	-0.691	0.489				
Residual variance intercept	0.095	0.033	2.876	0.004				
Gross-level interaction: effect on rand. slope (Math ach. t2 regressed on math ach. Q4 t1)								
Intercept random slope	-0.004	0.164	-0.023	0.982				
Math achievement t1	-0.184	0.168	-1.095	0.273				
Group ^{MAT}	0.143	0.154	0.928	0.353				
Group ^{MEET}	-0.154	0.147	-1.046	0.296				
Residual variance random slope	0.008	0.094	0.080	0.936				
Between (model 1b; ref = group^{MEET})								
Intercept math achievement t2	-0.219	0.701	-0.313	0.755				
Math achievement t1	0.928	0.167	5.565	0.000				
Group ^{MAT}	0.314	0.139	2.254	0.024				
Group ^{CONTROL}	0.083	0.120	0.691	0.489				
Residual variance intercept	0.095	0.033	2.876	0.004				
Gross-level interaction: effects on rand. slope (Math ach. t2 regressed on math ach. Q4 t1)								
Intercept random slope	-0.157	0.124	-1.267	0.205				
Math achievement t1	-0.184	0.168	-1.095	0.273				
Group ^{MAT}	0.296	0.106	2.791	0.005				
Group ^{CONTROL}	0.154	0.147	1.046	0.296				
Residual variance random slope	0.008	0.094	0.080	0.936				

Notes: Non-standardized regression coefficients with standard errors based on MLR estimation
 Latent level-specific decomposition of mathematics score t1 and t2

The level-1 effects remained unchanged. In terms of fixed effects at level 2, when this alternative random slope was introduced the only significant difference found was between classes in group^{MAT} and group^{MEET} ($b = 0.314$, $t = 2.254$, $p = .024$), just as in models 1 and 2. This means that in models 3a and 3b the mean gain in math achievement is higher in classes in group^{MAT} than in classes in group^{MEET}. The difference in mathematical achievement gain on class level between the classes in group^{MAT} and group^{CONTROL} is not significant ($b = 0.231$, $t = 1.700$, $p = .089$), which was also the case in model 1a with $p = .06$.

In the random part of models 3a and 3b the random slope coefficient was not dependent on the level of math achievement at t1. But, in model 3b the cross-level interaction for the contrast between group^{MAT} and group^{MEET} was significant ($b = 0.296$, $t = 2.791$, $p = .005$). This means that the students in the top quartile of math achievement in each class in group^{MAT} showed higher mathematical achievement gains than the top quartile of students in each class of group^{MEET}.

Discussion

This study aimed to analyze the null and unexpected effects of an intervention study that was designed to support low achievers in mathematics in an inclusive setting (Pfister, Stöckli, et al., 2015; Stöckli et al., 2014; Stöckli, 2019). Specifically, it looked at whether the mathematical achievement gain of low achievers in mathematics (RQ1) and students with high mathematical achievement (RQ 2) depended on their mathematical achievement level at the beginning of the study.

As expected, individual variables (IQ, SEN^{MATH}, gender, German comprehension) significantly influenced the residual change score in math achievement between t1 and t2 at the individual level. We had predicted that low achieving students in group^{MEET},

where teachers attended two in-service meetings to discuss the implementation of the intervention, would benefit more than those in the other groups. Indeed, the analyses of Pfister, Moser Opitz and Pauli (2015) indicated a tendency for an increased use of student support strategies by teachers in group^{MEET}. But the results of model 2a and 2b showed no significant effect for the students in the first quartile of math achievement in each class in group^{MEET} compared to students in the same quartile in the other two experimental conditions (group^{CONTROL} and group^{MAT}). The theoretical advantage conferred by treatment MEET was not confirmed, even for the main target group. Instead, models 3a and 3b showed that there was a positive effect for those students in the group^{MAT} classes (where teachers received the material) who already had a high level of math achievement (top quartile of each class) compared to the high achievers of each class in group^{MEET}.

Although the results show that the intervention did not have the desired effect, they do provide important insights for planning future studies and on how to improve the teaching of low achievers in mathematics (Jacob et al., 2019). First, there is no data-based explanation for the different outcomes for group^{MAT} and group^{MEET} and for the positive effect of the intervention on the high achieving students in the group^{MAT}. Classroom composition does not explain the differences (Stöckli, 2019). This suggests that other variables, which have not been collected, or not collected systematically, might have affected the results. In large samples, if the sampling process was unbiased, we would expect such variables to be neutralized. In smaller samples, like this one, such variables can play a significant role. Variables which could have affected the results include the number of students per class who have behavioral problems, the classroom management skills of the general education teacher (Farmer et al., 2019), his or her professional mathematical knowledge (Hill et al., 2005), the regu-

lar presence of a special education teacher (SET) in the classroom (Bottge et al., 2017), and the number of hours a SET was present in the classroom (Moser Opitz, Schnepel, et al., 2020). As reported by Pfister, Moser Opitz, and Pauli (2015), the results of the video analyses showed that a second teacher in the role of a SET was involved in mathematics lessons in some classes although the support did vary (support of a single student, co-teaching, mixed). Also, the quality of student support (scaffolding) offered by the teachers, both general education teacher and SET, varied a great deal. This underlines a fundamental challenge of research in complex inclusive education settings. Jones and Brownell (2014) characterize these settings as “nested instruction” and emphasize that it is difficult to disentangle the influence of the single teacher when different teachers work in different settings for varying lengths of time with different students. It is important that researchers develop ways to accommodate these challenges, although controlling for all of the relevant variables or even conducting studies with systematic variation would require very large sample sizes. While this may be realistic for survey studies, it is very challenging for intervention programs, which require a high level of commitment and engagement from participating teachers.

Because, contrary to the study objective, low achievers in mathematics did not benefit from this intervention - high achievers in group^{MAT} were the greatest beneficiaries - we must consider whether targeted support in an inclusive setting can meet the needs of low achieving students. Low achieving students need more individual, intensive support (Chodura et al., 2015; Ise et al., 2012). In another intervention study by Moser Opitz et al. (2017), which succeeded in improving the achievement of low achievers in mathematics, the program provided more individual support; one-to-one work was partially integrated into the regular classroom setting. A successful inclusive intervention program that focused on percent-

ages, by Kuhl et al. (2021), also included much tailored, individual work. These studies used a more individualized approach with differentiated diagnostic tasks than was the case in this study, although it should be noted that they were both conducted in secondary schools. It might be that inclusive programs are more suitable for older students whose self-regulated competence may be higher. We conclude that this intervention was not sufficiently tailored for the needs of low achievers or individualized.

Limitations

Perhaps conducting additional analyses using data from a study that did not produce satisfactory results is not advisable; it can be seen as an attempt to retrospectively “bend the data into the shape” or gloss over null effects or unwanted findings. However, here the results of the new analyses do not contradict the earlier findings. They help to clarify the reasons behind the outcome. Building groups based on a measurement at a given point in time may lead to pushing the regression towards the mean because the measurement error is asymmetrically dependent on the distance between each value and the mean. This is a minor concern in this context: The reliability of the mathematical achievement measure is high (.84) and the potential bias is the same for all three treatment conditions. Therefore, any possible regression effect should not have resulted in a false conclusion about the assessment of potential differential effects. Another limitation of the study is its focus on only the general education teacher with no data collected on cooperation with a SET. Finally, the design of the program allowed teachers to decide which topic (e.g. place value, number line) should be taught to which students and there were no concrete diagnostic tasks to help them make a determination. For future programs, it would be important to offer diagnostic tasks along with information on how to support low achievers in mathematics.

Conclusion

The results of this study contradict the accepted view that specific support in an inclusive setting is the best way to attain improvements in achievement of low achievers (e.g. Hinz & Köpfer 2016; Feuser 1989; Scherer et al. 2016). Classroom support alone is apparently not enough to address the deficits in the mathematical understanding of low achievers although being taken outside the classroom for special support on a regular basis affects the social acceptance of those students by their peers (Wiener & Tardif, 2004). While studies have shown that low achievers can modulate any rejection with cooperative behavior (Schnepel et al., 2021), for this to happen the students have to be able to participate in joint learning situations. It may therefore be that the social needs of low achievers in inclusive classrooms are in conflict with their academic needs. They need opportunities to cooperate with their peers to increase their social acceptance and they need individual support that must sometimes be provided outside of the classroom. The solution is to use multiple approaches, including group learning in an inclusive classroom and individualized measures (Moser Opitz et al., 2018). More research, which considers the complex topic of nested instruction (Jones & Brownell, 2014), is needed to develop and evaluate such settings.

References


- Andersson, U. (2010). Skill development in different components of arithmetic and basic cognitive functions: Findings from a 3-year longitudinal study of children with different types of learning difficulties. *Journal of Educational Psychology, 102*(1), 115–134. <https://doi.org/10.1037/a0016838>
- Bottge, B.A., Cohen, A.S., & Choi, H.-J. (2017). Comparisons of mathematics intervention effects in resource and inclusive classrooms. *Exceptional Children, 84*(2), 197–212. <https://doi.org/10.1177/0014402917736854>
- Cattell, R.B., Weiß, R.H., & Osterland, J. (1997). *Grundintelligenztest Skala 1: CFT 1. Test* (5., revidierte Aufl.). Hogrefe.
- Chan, W.W.L., Au, T.K., Lau, N.T.T., & Tang, J. (2017). Counting errors as a window onto children's place-value concept. *Contemporary Educational Psychology, 51*, 123–130. <https://doi.org/10.1016/j.cedpsych.2017.07.001>
- Chodura, S., Kuhn, J.T., & Holling, H. (2015). Interventions for children with mathematical difficulties: A meta-analysis. *Zeitschrift für Psychologie, 223*(2), 129–144. <https://doi.org/10.1027/2151-2604/a000211>
- Conlin, L.D., Kuo, E., & Hallinen, N.R. (2019). How null results can be significant for physics education research. *Physical Review Physics Education Research, 15*(2), 1-13. <https://doi.org/10.1103/PhysRevPhysEducRes.15.020104>
- Desoete, A., Ceulemans, A., Roeyers, H., & Huylebroeck, A. (2009). Subitizing or counting as possible screening variables for learning disabilities in mathematics education or learning? *Educational Research Review, 4*(1), 55–66. <https://doi.org/10.1016/j.edurev.2008.11.003>


- Farmer, T.W., Hamm, J.V., Dawes, M., Barco-Alva, K., & Cross, J.R. (2019). Promoting inclusive communities in diverse classrooms: Teacher attunement and social dynamics management. *Educational Psychology, 54*(4), 286–305. <https://doi.org/10.1080/00461520.2019.1635020>
- Feuser, G. (1989). Allgemeine integrative Pädagogik und entwicklungslogische Didaktik. *Behindertenpädagogik, 28*(1), 4–48. <https://doi.org/10.25656/01:17007>
- Freeseemann, O. (2014). Schwache Rechnerinnen und Rechner fördern. Eine Interventionsstudie an Haupt-, Gesamt und Förderschulen. Springer Spektrum.
- Fuchs, L.S., Fuchs, D., Compton, D.L., Wehby, J., Schumacher, R.F., Gersten, R., & Jordan, N.C. (2015). Inclusion versus specialized intervention for very-low-performing students: What does access mean in an era of academic challenge? *Exceptional Children, 81*(2), 134–157. <https://doi.org/10.1177/0014402914551743>
- Gage, N.A., Cook, B., & Reichow, B. (2017). Publication bias in special education meta-analyses. *Exceptional Children, 83*(4), 428–445. <https://doi.org/10.1177/0014402917691016>
- Gersten, R., Chard, D.J., Jayanthi, M., Baker, S.K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research, 79*(3), 1202–1242. <https://doi.org/10.3102/0034654309334431>
- Hagermoser Sanetti, L.M., & Fallon, L.M. (2011). Treatment integrity assessment: How estimates of adherence, quality, and exposure influence interpretation of implementation. *Journal of Educational & Psychological Consultation, 21*(3), 209–232. <https://doi.org/10.1080/10474412.2011.595163>
- Hill, H.C., Rowan, B., & Ball, D.L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371–406. <https://doi.org/10.3102/00028312042002371>
- Hinz, A., & Köpfer, A. (2016). Unterstützung trotz Dekategorisierung? Beispiele für Unterstützung durch Dekategorisierung. *VHN, 85*(1), 36–47. <https://doi.org/10.2378/vhn2016.art04d>
- Ise, E., Dolle, K., Pixner, S., & Schulte-Körne, G. (2012). Effektive Förderung rechenschwacher Kinder. Eine Metaanalyse. *Kindheit und Entwicklung, 21*(3), 181–192. <https://doi.org/10.1026/0942-5403/a000083>
- Jacob R.T., Doolittle F., Kemple J., & Somers M-A. (2019). A framework for learning from null results. *Educational Researcher, 48*(9), 580–589. <https://doi.org/10.3102/0013189X19891955>
- Jitendra A.K., Alghamdi, A., Edmunds, R., McKeveit, N.M., Mouanoutoua, J., & Roesslein R. (2021). The effects of tier 2 mathematics interventions for students with mathematics difficulties: A meta-analysis. *Exceptional Children, 87*(3), 307–325. <https://doi.org/10.1177/0014402920969187>
- Jitendra, A.K., Harwell, M.R., Dupuis, D.N., & Karl, S.R. (2017). A randomized trial of the effects of schema-based instruction on proportional problem-solving for students with mathematics problem-solving difficulties. *Journal of Learning Disabilities, 50*(3), 322–336. <https://doi.org/10.1177/0022219416629646>
- Jones, N.D., & Brownell, M.T. (2014). Examining the use of classroom observations in the evaluation of special education teachers. *Assessment for Effective Intervention, 39*(2), 112–124. <https://doi.org/10.1177/1534508413514103>
- Kuhl, J., Prediger, S., Schulze, S., & Wittich, C. (2021). Inklusiver Mathematikunterricht in der Sekundarstufe – Eine Pilotstudie zur Prozentrechnung. *Unterrichtswissenschaft. https://doi.org/10.1007/s42010-021-00125-8*
- Lipowsky, F. (2004). Was macht Fortbildungen für Lehrkräfte erfolgreich? Befunde der Forschung und mögliche Konsequenzen für die Praxis. *Die Deutsche Schule 96*(4), 462–479.


- Lüdtke, O., Marsh, H.W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*(3), 203–229. <https://doi.org/10.1037/a0012869>
- Moser Opitz, E. (2013). Rechenschwäche/ Dyskalkulie. Theoretische Klärungen und empirische Studien an betroffenen Schülerinnen und Schülern. 2. Aufl. Haupt.
- Moser Opitz, E., Schnepel S., Krähenmann, H., Jandl, S. Felder, F. & Sermier Dessemontet, R. (2020). The impact of special education resources and the general and the special education teacher's competence on pupil mathematical achievement gain in inclusive classrooms. *International Journal of Inclusive Education*. <https://doi.org/10.1080/13603116.2020.1821451>
- Moser Opitz, E., Freeseemann, O., Prediger, S., Grob, U., Matull, I., & Hußmann, S. (2017). Remediation for students with mathematics difficulties: An intervention study in middle schools. *Journal of Learning Disabilities, 50*(6), 724–736. <https://doi.org/10.1177/0022219416668323>
- Moser Opitz, E., Stöckli, M., Grob, U., Nührenbörger, M., & Reusser, L. (2020). *BASIS-MATH-G 2+. Gruppentest zur Basisdiagnostik Mathematik für das vierte Quartal der 2. Klasse und das erste Quartal der 3. Klasse*. Hogrefe.
- Moser Opitz, E., Stöckli, M., Grob, U., Nührenbörger, M., & Reusser, L. (2019). *BASIS-MATH-G 3+. Gruppentest zur Basisdiagnostik Mathematik für das vierte Quartal der 3. Klasse und das erste Quartal der 4. Klasse*. Hogrefe.
- Moser Opitz, E., Grob, U., Wittich, C., Häsel-Weide, U. & Nührenbörger, M. (2018). Fostering the computation competence of low achievers through cooperative learning in inclusive classrooms: A longitudinal study. *Learning Disabilities: A Contemporary Journal, 16*(1), 19–35. <https://eric.ed.gov/?id=EJ1179942>
- Muthén, L.K. & Muthén, B.O. (2017). *Mplus User's Guide. Eighth Edition*. Muthén & Muthén.
- Paulus, C. (2009). Die «Bücheraufgabe» zur Bestimmung des kulturellen Kapitals bei Grundschulern. Retrieved from <http://hdl.handle.net/20.500.11780/3344>
- Pfister, M., Stöckli, M., Moser Opitz, E., & Pauli, C. (2015). Inklusiven Mathematikunterricht erforschen: Herausforderungen und erste Ergebnisse aus einer Längsschnittstudie. *Unterrichtswissenschaft 43*(1), 53–67.
- Pfister, M., Moser Opitz, E., & Pauli, C. (2015). Scaffolding for mathematics teaching in inclusive primary classrooms: A video study. *ZDM Mathematics Education, 47*(7), 1079–1092. <https://doi.org/10.1007/s11858-015-0713-4>
- Powell, S.R., Cirino, P.T., & Malone, A.S. (2017). Child-level predictors of responsiveness to evidence-based mathematics intervention. *Exceptional Children, 83*(4), 359–377. <https://doi.org/10.1177/0014402917690728>
- Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three basic dimensions. *ZDM Mathematics Education, 50*(3), 407–426. <https://doi.org/10.1007/s11858-018-0918-4>
- Raudenbush, S.W., Bryk, A.S., & Congdon, R.T. (2013). *HLM 7.01 for Windows* [Computer software]. Scientific Software International, Inc.
- Scherer, P., Beswick, K., DeBlois, L., Healy, L., & Moser Opitz, E. (2016). Assistance of students with mathematical learning difficulties: How can research support practice? *ZDM Mathematics Education, 48*, 633–649. <https://doi.org/10.1007/s11858-016-0800-1>
- Schnepel, S., Garrote, A., & Moser Opitz, E. (2021). Disentangling the relationship between mathematical achievement, social status, and social skills in inclusive classrooms. *Empirische Sonderpädagogik, 13*(2), 148–166. <https://doi.org/10.25656/01:23576>

- Stevens, E.A., Rodgers, M.A., & Powell, S.R. (2018). Mathematics interventions for upper elementary and secondary students: A meta-analysis of research. *Remedial and Special Education, 39*(6), 327–340. <https://doi.org/10.1177/0741932517731887>
- Stöckli, M. (2019). *Unterrichtsintegrierte Förderung im Mathematikunterricht: Eine empirische Studie in der Primarschule* [Dissertation, Universität Zürich]. Zurich Open Repository and Archive. <https://doi.org/10.5167/uzh-177335>
- Stöckli, M., Moser Opitz, E., Pfister, M., & Reusser, L. (2014). Gezielt fördern, differenzieren und trotzdem gemeinsam lernen. Überlegungen zum inklusiven Mathematikunterricht. *Sonderpädagogische Förderung heute, 59*(1), 44–56.
- Wiener, J., & Tardif, C.Y. (2004). Social and emotional functioning of students with learning disabilities: Does special education placement make a difference? *Learning Disabilities Research & Practice, 19*(1), 20–32. <https://doi.org/10.1111/j.1540-5826.2004.00086.x>
- Wood, T., Mazzocco, M.M.M., Calhoon, M.B., Coyne Crowe, E., & McDonald Connor, C. (2020). The effect of peer-assisted mathematics learning opportunities in first grade classrooms: What works for whom? *Journal of Research on Educational Effectiveness, 13*(4), 601–624. <https://doi.org/10.1080/19345747.2020.1772422>
- Zhang, D., & Xin, Y.P. (2012). A follow-up meta-analysis for word-problem-solving interventions for students with mathematics difficulties. *The Journal of Educational Research, 105*(5), 303–318. <https://doi.org/10.1080/00220671.2011.627397>
- Zurbruggen, C. (2016). *Schulklasseneffekte. Schülerinnen und Schüler zwischen komparativen und normativen Einflüssen*. Springer.

Author Note:

 Urs Grob
<https://orcid.org/0000-0002-6671-7085>

 Elisabeth Moser Opitz
<https://orcid.org/0000-0002-5243-4770>

 Meret Stöckli
<https://orcid.org/0000-0003-1518-5960>

All authors were equal contributors to this paper.

This work was supported by the Swiss National Science Foundation (grant number 134652)

Corresponding author:

Elisabeth Moser Opitz
 Universität Zürich
 Institut für Erziehungswissenschaft
 Freiestrasse 36
 CH-8032 Zürich
elisabeth.moseropitz@uzh.ch

Erstmals eingereicht: 10.01.2022

Überarbeitung eingereicht: 09.04.2022

Angenommen: 15.07.2022