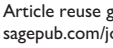
Research Study

Hammill Institute on Disabilities

Remedial and Special Education
1–11

DOI: 10.1177/07419325221102537
rase.sagepub.com


Improving Numeracy Skills in First Graders with low performance in early numeracy: A Randomized Controlled Trial

Anita Lopez-Pedersen, PhD[1], Riikka Mononen, PhD[1,2], Pirjo Aunio, PhD[3], Ronny Scherer, PhD[1], and Monica Melby-Lervåg, PhD[1]

Abstract

Children with low performance in early numeracy are at risk of facing learning difficulties in mathematics, but few trials have examined how this can be ameliorated. A total of 120 first-grade children (M_{age} = 6.4 years) were randomly assigned to an intervention or a control condition. The 14-week intervention targeted early numeracy skills and was delivered in small groups three times a week. Immediately after the initial 8-week intervention phase, moderate and positive effects were found on early numeracy (d = 0.19), word problem solving (d = 0.41), and approximate number sense (d = 0.35). However, only the effects on word problems were significant, and all effects disappeared after the children undertook a second 6-week intervention phase. Overall, results indicate that (a) early numeracy skills are malleable in low-performing children, but (b) frequent and long-term interventions are needed for the positive effects to last.

Keywords

early numeracy, numeracy intervention, randomized controlled trial, structural equation modeling

Learning difficulties in mathematics are frequent and a common cause of special education. However, few studies have empirically examined how to prevent and ameliorate such difficulties (Chodura et al., 2015; Dennis et al., 2016). Here, we present a randomized controlled trial of a numeracy intervention for first graders with low performance in numeracy skills. Early numeracy depends on mastering a range of skills, for example, comparing magnitudes (approximate number sense), counting, knowing number symbols, recognizing (un)structured quantities, and estimating quantities (Gersten et al., 2012; Moeller et al., 2011). Longitudinal studies have shown that early numeracy creates a basis for children's learning trajectories to school mathematics (Duncan et al., 2007; Jordan et al., 2009; Nguyen et al., 2016). Yet, children's levels of early numeracy vary widely even before formal schooling, and the differences in performance levels often persist after (Aunola et al., 2004; Seethaler & Fuchs, 2011).

Previous Numeracy Interventions in Kindergarten and Early Grades

There are a number of meta-analyses of interventions targeting mathematical learning difficulties (Chodura et al., 2015; Dennis et al., 2016; Gersten et al., 2009; Jitendra et al., 2018; Kroesbergen & Van Luit, 2003; Monei & Pedro, 2017; Wang et al., 2016). Chodura et al. (2015), whose inclusion criteria overlap with the sample in our study, have examined the effects of interventions in 6- to 12-year olds with mathematical difficulties in pre/post control group studies. Based on 35 studies overall, there were large intervention gains in number skills and arithmetic (d = 0.83). Furthermore, moderator analyses have shown that the most efficient interventions were based on direct and assisted teaching. However, most studies had poor or no randomization and low power (>30 in each group).

In another meta-analysis, Dennis et al. (2016) also examined children with mathematical difficulties in pre/post control group studies. Their results showed promising findings for interventions that included peer-assisted training and explicit instructions in small groups, but overall effects were only moderate (d = 0.55, k = 25). The reason for the discrepancy in mean effect size between these two meta-analyses may lie in the selection criteria, as Dennis et al. (2016) used a more lenient definition of mathematical

[1]University of Oslo, Norway
[2]University of Oulu, Finland
[3]University of Helsinki, Finland

Corresponding Author:
Anita Lopez-Pedersen, Department of Special Needs Education, University of Oslo, Sem Sælands vei 7, 0371 Oslo, Norway.
Email: anita.lopez-pedersen@oslomet.no

Associate Editor: Mian Wang

difficulties compared to Chodura et al. (2015). Ultimately, both meta-analyses show that mathematics interventions can be effective, but that moderators may be important to the size of the effects (Dennis et al., 2016).

If we examine the randomized trials included in these reviews more closely, they show mixed effects. For instance, Fuchs et al. (2013) have examined the effects of strategic number knowledge intervention for first graders in three groups: a control group, a group with strategic counting *without* speeded practice (i.e., reinforcing thoughtful application to support reasoning of strategies during fact retrieval), and a group with strategic counting *and* speeded practice (i.e., promoting quick response to support fact retrieval). This intervention lasted for 16 weeks (30-minute sessions, three times weekly). Strategic counting *without* speeded practice was seen to improve number combination fluency compared to the control condition on immediate posttest ($d = 0.43$), while strategic counting *with* speeded practice improved number combination fluency and transfer to procedural calculations compared to both competing conditions on immediate posttest ($d = 0.67$).

Gersten et al. (2015) conducted a scale-up trial of the "Number Rockets" (Fuchs et al., 2005) intervention program, a small-group intervention for at-risk first graders focusing on number operations (30 hours of small-group work). Children in the intervention group outperformed the control group on a broad measure of mathematics proficiency on the immediate posttest ($d = 0.34$).

Clarke et al. (2016) have examined the efficacy of a kindergarten early numeracy intervention program ("ROOTS"), focusing on developing whole-number understanding for children assessed as at-risk in mathematics (50 20-minute sessions over 10 weeks). Results across 29 classrooms showed that children in the intervention group outperformed the control group ($d = 0.28$ for oral counting, $d = 0.75$ for early numeracy, and $d = 0.48$ for early number sense). However, no effects were found in the follow-up posttest, indicating that the initial positive impacts of the intervention did not remain long-term. Hence, Clarke et al. (2016) raised the concern of limited impact on long-term achievement in mathematics interventions. Notably, effects from ROOTS were replicated by Doabler et al. (2016) with similar findings (effect sizes ranging from $d = 0.31$ to 1.08).

A consistent finding across studies is that effects fade out after a seemingly effective mathematics intervention has ended (Bailey, 2019). Little is known about the nature of fade-out effects and their influencing factors in the context of randomized controlled trial (RCT) studies. Two hypotheses have been suggested: first, the constraining content hypothesis, which suggests that fade-out effects are due to environmental factors, given that subsequent instruction does not build on the skills learned during the intervention. Second, the preexisting differences hypothesis (Bailey et al., 2016) suggests that fade out is due to stable, underlying characteristics in mathematics that cause children to revert to their previous individual trajectories (Bailey et al., 2016).

Research Aim and Questions

Evidently, previous studies have identified promising effects immediately after intervention, but these effects fade out as soon as the intervention is taken away. Here, we present a study that will contribute to knowledge in this area. First, we examine effects from an early numeracy intervention in a developmental period in which numeracy skills are presumed to be malleable. Second, the present study attempts to prevent fade-out effects by adding a second intervention phase that serves as a refresher of the intervention content.

Accordingly, we aim to respond to the following research questions:

1. Does an early numeracy intervention lead to pretest/posttest differences between treatment and control groups in early numeracy, word problem solving, arithmetic skills, and approximate number sense (*immediate intervention effects*)?
2. Does including a second intervention phase lead to pre/follow up-test differences in outcomes between treatment and control groups (*follow-up intervention effects*)?

Method

Participants

All children born in 2010 and attending first grade in two municipalities in Norway were invited to participate in the study. Children start school at the age of six in Norway; this resulted in 369 initial participants. The CONSORT diagram in Figure A1 in the Supplemental Appendix depicts the flow of participants throughout the study (Schulz et al., 2010). Ethical approval was obtained from the Norwegian Social Science Data Services, and informed parental consent was given.

The children were selected based on a screening with the Early Numeracy Screener (Lopez-Pedersen et al., 2021), consisting of 52 items measuring early numeracy skills, understanding numerical relational skills, counting skills, and basic arithmetic skills. The tasks in the screening measure are like those assessed by other early numeracy measures (Clements et al., 2008; Jordan et al., 2007). The reliability of the screening measure in our sample was Cronbach's $\alpha = .943$. We identified 32% of the children with the lowest scores in the early numeracy screener ($n = 120$, 57% girls) for further participation in the study ($M_{age} = 77$ months, $SD = 3.94$ months).

Table 1. Means, Standard Deviations, Cronbach's α, and Effect Sizes for All Measures at All Time Point in the Intervention Group and the Control Group.

		M (*SD*)		
Measures	Time point	Intervention group	Control group	Cronbach's α
Relational skills	Pretest	14.00 (3.33)	15.25 (3.66)	.859
	Posttest 1	30.49 (6.87)	30.30 (7.80)	.658
	Posttest 2	34.74 (6.74)	35.82 (6.79)	.889
	Follow-up test	20.71 (5.05)	23.33 (4.60)	.820
Counting skills	Pretest	21.58 (6.24)	23.62 (7.15)	.722
	Posttest 1	16.51 (2.89)	17.35 (3.52)	.877
	Posttest 2	17.72 (3.51)	18.57 (3.72)	.799
	Follow-up test	21.80 (7.34)	22.53 (6.81)	.918
Word problems TM	Pretest	1.57 (1.38)	2.10 (1.74)	.576
	Posttest 1	3.03 (1.64)	3.10 (1.83)	.581
	Posttest 2	3.66 (2.12)	4.33 (1.88)	.681
	Follow-up test	4.55 (2.07)	4.83 (1.97)	.683
Word problems W	Pretest	8.98 (2.70)	9.87 (2.83)	.727
	Posttest 1	11.54 (2.96)	9.87 (2.83)	.777
	Posttest 2	12.48 (3.82)	12.78 (3.79)	.828
	Follow-up test	13.95 (4.07)	14.29 (3.64)	.810
Dot comparison	Pretest	9.02 (2.94)	8.98 (2.38)	.677
	Posttest 1	10.51 (3.18)	10.28 (3.02)	.787
	Posttest 2	13.29 (4.05)	12.52 (3.66)	.813
	Follow-up test	15.47 (4.09)	15.72 (4.44)	.832
Digit comparison	Pretest	13.40 (4.04)	14.50 (3.33)	.840
	Posttest 1	16.80 (3.60)	16.22 (4.12)	.858
	Posttest 2	17.53 (4.55)	17.48 (3.39)	.876
	Follow-up test	19.82 (4.17)	18.88 (4.81)	.892
Addition	Pretest	3.78 (3.24)	4.82 (3.09)	.888
	Posttest 1	6.92 (2.44)	7.30 (2.42)	.807
	Posttest 2	7.69 (2.79)	8.33 (1.98)	.873
	Follow-up test	6.78 (2.64)	7.10 (2.26)	.802
Subtraction	Pretest	0.45 (1.19)	0.83 (2.01)	.872
	Posttest 1	5.14 (3.06)	5.25 (3.35)	.886
	Posttest 2	6.00 (3.29)	7.05 (2.57)	.868
	Follow-up test	7.53 (2.36)	7.67 (2.66)	.827

Note. Word problems TM = Word problem items from the research-developed early numeracy test, Word problems W = Word problem items from WISC-IV. Measures of counting and relational skills in t4 was changed in order to avoid ceiling effects. Items that 95% of the children correctly solved on t3 were taken out and the remaing items with increased difficulty level. Thus lower means is due to fewer items.

Two of the authors randomized the children at the individual level by using random.org (https://www.random.org); using the same program, we applied blocking to ensure an equal number of children in both groups. The study had little attrition: only 5.8% ($n = 7$) of participants dropped out due to moving school districts. Little's MCAR (Missing Completely at Random) Test (R. J. A. Little, 1988) of the pretest data indicated that the data were likely to follow a missing-completely-at-random mechanism rather than a missing-at-random mechanism, $\chi^2(105) = 79.0$, $p = .97$. We therefore performed the full-information maximum-likelihood procedure to handle the missing data (Enders, 2010).

Measures

Children were assessed individually at preintervention (t1), at immediate posttest at the end of the first 8 weeks of intervention (t2), at immediate second posttest after receiving the secondary intervention once a week for 6 weeks (t3), and at follow-up 6 months after the intervention ended (t4). All testing was conducted by trained research assistants in the children's schools during school time. Internal consistencies of all measures were satisfactory (see Table 1).

Early numeracy skills. Early numeracy skills were assessed using items of counting and numerical relational skills from

a test custom-developed for this study (Aunio et al., 2016). This test consisted of 24 counting tasks, measuring number-quantity correspondence, enumeration, and number sequences. The 24 items measured numerical relational skills, such as comparing numbers; for example, identifying the smallest/largest number within the number range 1 to 201 (e.g., 22–19–28), identifying quantities with instructions such as "one more than," "one less than," and items measuring ordinal numbers without time limit. Each item was given one point for the correct answer and zero for the incorrect answer.

Word problem solving. Word problem solving was assessed using items from two tests: WISC-IV arithmetic tasks (Wechsler et al., 2003) and a custom-developed test for this study (Aunio et al., 2016). With the former, the WISC-IV arithmetic tasks contained 34 arithmetic word problems, with a time limit of 30 seconds per item and a stopping rule of four consecutive errors. With the latter, word problem solving was assessed using an 8-item test. In both tests, the children were given word problems (read aloud to them) and then asked to solve them mentally, reporting their answers to the assessor. Each item was given one point for the correct answer and zero for the incorrect answer.

Arithmetic skills. Arithmetic skills were assessed by measuring addition and subtraction skills using items from a test developed for this study (Aunio et al., 2016). The children were asked to perform 10 addition tasks (using paper and pencil) without a time limit. Eight of the tasks were in the number range of 0 to 20, and two were in the range of 10 to 30. For the subtraction items, the children were asked to perform 10 subtraction tasks (paper and pencil) without a time limit. All tasks were in the number range of 0 to 20. Each item was given one point for the correct answer and zero for the incorrect answer.

Approximate number sense (ANS). Approximate number sense (ANS) was assessed by measuring dots and digit comparison skills using two tasks from the Test of Basic Arithmetic and Numeracy Skills (Brigstocke et al., 2016). For the dot comparison tasks, the test presented arrays of dots randomly arranged within a 2.5 cm^2 box on a white background. A series of items with two adjacent boxes were given, and the children were asked to quickly tick the box with the largest number of dots and to complete as many boxes as possible within 30 seconds. The digit comparison tasks were presented in columns of two digits next to each other, and the children were asked to mark the larger of the two numbers. The children completed as many tasks as they could within 30 seconds and were given one point for each correct answer and zero for each wrong answer.

Intervention Program

When designing interventions for early numeracy skills, targeted skills generally fall into three domains: understanding numerical relations, counting skills, and basic arithmetic skills (e.g., Aunio & Räsänen, 2016; Jordan et al., 2009; Purpura et al., 2013). The present intervention program is focused on counting in the number range 1 to 20 and is based on the model of development of core numeracy skills theorized by Aunio and Räsänen (2016). Explicit teaching serves as an instructional feature because it has repeatedly been proven to be an effective approach (e.g., Chodura et al., 2015; Kroesbergen & Van Luit, 2003). (See Tables A1 and A2 in the online supplemental materials for the detailed content of each intervention session).

The intervention sessions were conducted in small pull-out groups consisting of four to six children and started with a short warm-up activity related to the content to be taught or with brief repetition of skills practiced in the previous session. A teacher-led activity followed, including modeling of new concepts and strategies. This was followed by children working in pairs, with hands-on activities (e.g., games and using manipulatives) guided by the teacher. At the very end of each session, the children completed a short individual written task. After every two small-group sessions, each child attended one 15 to 20-minute individual session with the intervention teacher. The objective of this individual session was to give the teacher an opportunity to work even more closely with the children and to give the teacher additional insight into each child's learning trajectory in early numeracy learning.

Procedure

The intervention condition comprises two phases. The first phase was administered to the children three times a week for 8 weeks by trained teachers and special educational needs teachers at school. A total of 24 sessions (16 small-group and eight individual sessions) were delivered, amounting to approximately 130 minutes each week. The second phase of the intervention started 2 weeks after the first intervention phase had ended. This phase consisted of six instructional sessions, once a week, over a total of 6 weeks. Content-wise, the sessions in the second phase repeated those of the initial 8-week intervention. Each intervention teacher received the material prior to the start of intervention, as well as training and practice in using the material.

Treatment Fidelity

During the intervention, we monitored the implementation of the intervention program using audio recordings. In addition, we used logs to note the children's attendance.

A random selection of 10% of the sessions across all schools was checked, and at least one session per teacher was checked. These sessions demonstrated 100% consistency between the audio recordings and the events reported in the logs. There was little absence from the intervention: on average, the children completed 27 out of 30 intervention sessions, yielding an absence rate of 10 %.

Statistical Analyses

We performed structural equation modeling (SEM) because it has the advantage of being a flexible framework for integrating observed and unobserved variables and including multiple groups and time points (Kline, 2016). We specified models that represented the key constructs in our study as either manifest or latent variables. These models describe the intervention effects at two measurement points (after controlling for children's performance at t1) and allowed us to examine the immediate and follow-up effects of the intervention. To sustain acceptable power, we decided to include only three rather than all four measurement points in the analytic models and to estimate separate models for each construct. In fact, power analyses in the R package "WebPower" version 0.5.2 (Zhang & Yuan, 2018) indicated reasonable power to detect small but significant intervention effects (see Supplementary Material S3 https://osf.io/pb2zn/). Specifically, to test the immediate intervention effects, we chose the outcome variables at t2 and t3, given that these two measurement points were close to each other. To test the follow-up effects, we chose the outcome variables at t4.

All variables were standardized on the dependent variables; thus, the path coefficients can be interpreted as differences in *SD* units (Cohen's *d*). We used intention to treat (ITT) analyses that included all children who received the pretest, irrespective of how many sessions they had actually taken. We performed all analyses using the R packages "lavaan" version 0.6–6 (Rosseel, 2012), "semTools" version 0.5–3 (Jorgensen et al., 2020), and "semPlot" version 1.1.2 (Epskamp, 2015), utilizing maximum-likelihood estimation and treating missing data via the full-information maximum-likelihood procedure (Enders, 2010). Supplementary Materials S3 and S4 https://osf.io/pb2zn/ provide the respective syntax and output of the analyses in R.

The data presented in this study has a partially nested design; that is, while students who were assigned to the intervention condition received the intervention in small groups (four to six students per group, 12 groups in total), students in the control condition did not. Several univariate and multivariate approaches have been proposed to efficiently consider this data structure, such as multigroup multilevel structural equation modeling (e.g., Candlish et al., 2018; Sterba et al., 2014). Despite wanting to do so, we were unable to model the partially nested structure of the data explicitly; given the small and unbalanced sample sizes at the cluster level, the parameters derived from such models were unacceptably large and did not result in reliable estimates of the intervention effects.

To evaluate the fit of the structural equation models that contained latent variables, we considered the common guidelines for model fit. These guidelines suggest an acceptable fit to the data if the Comparative Fit Index (CFI) exceeds .95, the Root Mean Square Error of Approximation (RMSEA) is less than .08, and the Standardized Root Mean Square Residual (SRMR) is less than .10 (Hu & Bentler, 1999; Marsh et al., 2005). We further tested for the invariance of the measurement models over time by specifying a configural, a metric, and a scalar invariance model, and comparing them against each other (T. D. Little, 2013). The more constrained model could be retained if the CFI did not decrease by more than .010, the RMSEA did not increase by more than .015, and the SRMR did not increase by more than .030 after introducing the equality constraints on factor loadings and intercepts to the model, (see Khojasteh & Lo, 2015; Putnick & Bornstein, 2016).

Results

Descriptive Statistics and Correlations

Table 1 shows the means, standard deviations, and reliabilities for all measures. The distribution of the variables was acceptable, except for subtraction, which had a floor effect at pretest. Supplementary Material S1 exhibits the full correlation matrix of these variables.

Measurement invariance testing. Given that the indicators of the models of approximate number sense and word problem solving were closely related (see Supplementary Material S1 and S3, https://osf.io/pb2zn/), we represented these constructs as latent variables. Testing for the invariance of the measurement model, scalar invariance held between groups for the construct of *word problem solving*. Hence, group comparisons were not affected by the differential functioning of the indicators of the two constructs (Millsap, 2011), and sufficient comparability over time was evident for the structural equation modeling of the intervention effects. As for *approximate number sense* over time, we found support for metric invariance (see Supplementary Material S2, https://osf.io/pb2zn/). This finding applied to both sets of measurement occasions: t1, t2, and t3, and t1, t2, and t4. Scalar invariance held for both sets across the two groups (control vs. intervention group). Thus, for *word problem solving* and *approximate number sense*, we carried out the analyses using latent variables. For early *numeracy* and *arithmetic*, we did not obtain measurement invariance. Due to the lack of invariance and the poor fit of the models with latent variables, we carried out these analyses on the

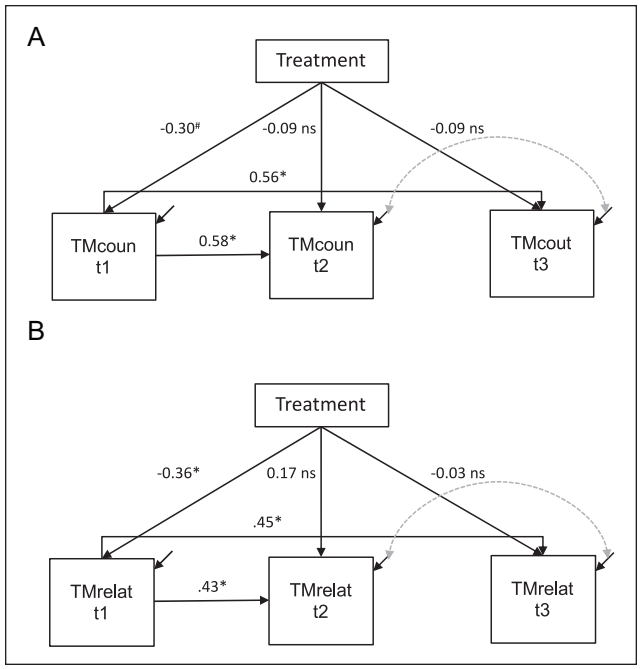


Figure 1. Effects of the intervention on early numeracy skills represented by (A) counting skills (TMcount1-3) and (B) numeric relations skills (TMrelatt1-3) at pretest, first posttest, and second posttest.
Note. STDY parameters shown. The variable treatment is binary (1 = *intervention group*, 0 = *control group*).
*$p < .05$. #$p < .10$, ns = statistically not significant ($p > .10$).

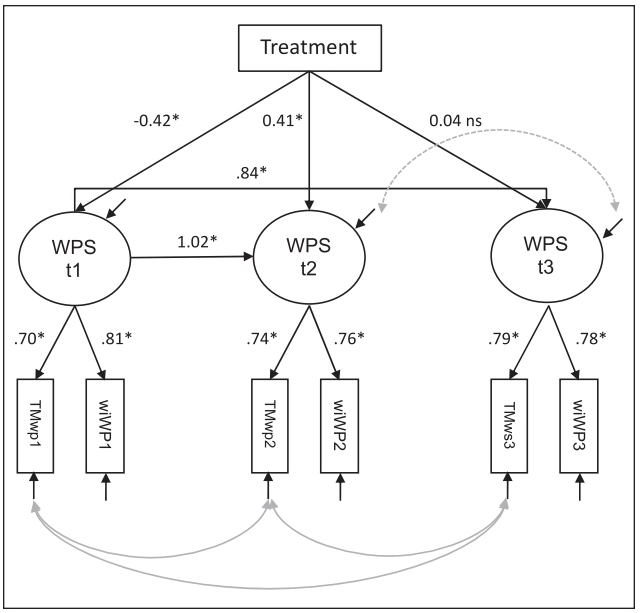


Figure 2. Effects of the intervention on word problem solving skills (WPSt1-3) at pretest, first posttest, and second posttest.
Note. STDY parameters shown. The variable treatment is binary (1 = *intervention group*, 0 = *control group*). WPS = word problem solving skills.
*$p < .05$, ns = statistically not significant ($p > .10$).

observed manifest variables only (for details, see Supplementary Material S1, https://osf.io/pb2zn/).

Structural Equation Modeling of the Intervention Effects

Early numeracy skills

First and second posttest. As mentioned, we used counting skills and numerical relations as manifest scores due to a lack of invariance and examined their effects in the intervention separately. The resulting models were exactly identified and had a perfect fit to the data. Figure 1 shows the effects on the models at posttest (immediately after 8 weeks of training three times per week) and at follow-up posttest (after an additional 6 weeks of training once a week). Overall, the intervention effects were small and insignificant (counting skills: $d = -0.09$, $p > .10$ for both t2 and t3; numerical relations: $d = 0.17$ for t2 and $d = -0.03$, $p > .10$ for t3).

Follow-up test 6 months after intervention. We specified a model like the one presented in Figure 1, except with t4 instead of t3. The model was exactly identified and thus had a perfect fit to the data. The treatment effects for t4 were not significant (counting skills: $d = 0.02$, $p = .90$; numerical relational skills: $d = -0.19$, $p = .23$).

Word problems

First and second posttest. Since we obtained invariance, the model for word problems consisted of one latent variable, with word problems from WISC-IV and from the test developed for the study as indicators. Figure 2 shows the effects of the intervention on word problems at the first posttest and second posttest. This model exhibited an excellent fit to the data: χ^2 (9) = 6.69, $p = .67$, RMSEA = .000 (90% CI = .000 - .198), CFI = 1.000, SRMR = .036. There was a significant and moderate effect at first posttest ($d = 0.41$, $p < .05$). However, when children received the sessions only once a week, the effects faded out at second posttest ($d = 0.04$, $p = .81$). Notably, after correcting for multiple corrections using the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995), results were no longer significant ($p = .15$).

Follow-up test 6 months after intervention. Once again, we found invariance and specified a model like the one presented in Figure 3, except with t4 instead of t3. This model exhibited an excellent fit to the data: χ^2 (8) = 5.11, $p = .75$, RMSEA = .000, 90% CI = .000–.077, CFI = 1.000, SRMR = .035. Considering the fade-out effect at second posttest, there were no significant effects at follow-up ($d = -0.41$, $p = .59$).

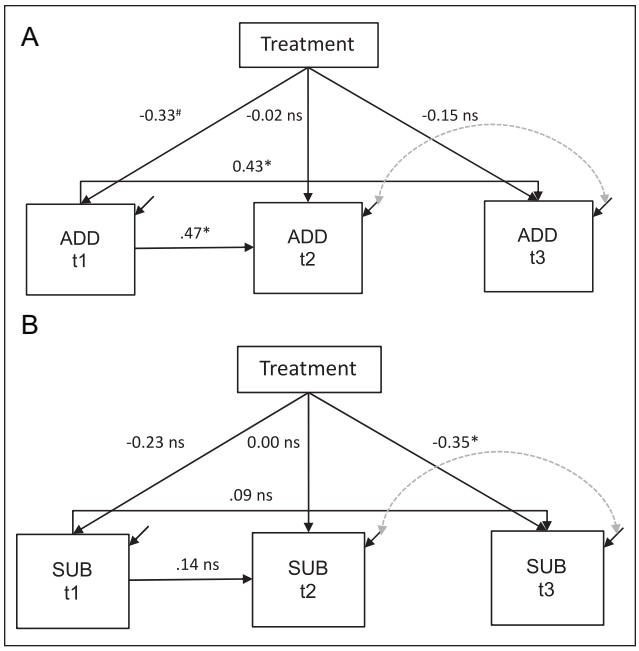


Figure 3. Effects of the intervention on arithmetic skills represented by (A) addition skills (ADDt1-3) and (B) subtraction skills (SUBt1-3) at pretest, first posttest, and second posttest.
Note. STDY parameters shown. The variable treatment is binary (1 = *intervention group*, 0 = *control group*).
*$p < .05$. #$p < .10$, ns = statistically not significant ($p > .10$).

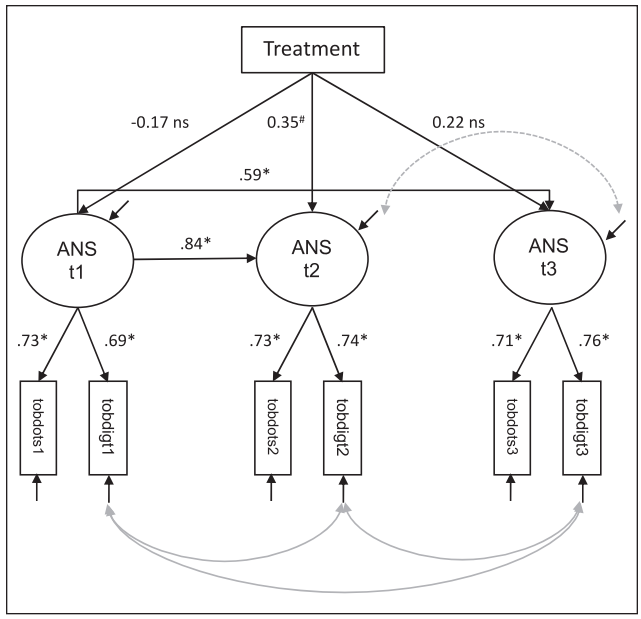


Figure 4. Effects of the intervention on approximate number sense (ANS) represented by children's performance on dot comparisons (tobdots1-3) and digit comparisons (tobdigt1-3) at pretest, first posttest, and second posttest.
Note. STDY parameters shown. The variable treatment is binary (1 = *intervention group*, 0 = *control group*). ANS = approximate number sense.
*$p < .05$. #$p < .10$, ns = statistically not significant ($p > .10$).

Arithmetic skills

First and second posttest. Due to a lack of measurement invariance and poor model fit, we compared the treatment effects over time for each of the two manifest indicators separately. The fit for the measurement model with all three measurement points was excellent: χ^2 (3) = 1.99, p = .57, RMSEA = .000 (90% CI = .000 - .132), CFI = 1.000, SRMR = .013. The resulting models and their parameters are shown in Figure 3. Overall, the intervention effects at t2 were insignificant (addition: $d = -0.02$, $p > .10$; subtraction: $d = 0.00$, $p > .10$). Fade-out effects occurred at t3 (addition: $d = -0.15$, $p > .10$; subtraction: $d = -0.35$, $p < .05$).

Follow-up test 6 months after intervention. As with the measurement model for t1 to t3, the measurement model for t1, t2, and t4 exhibited an excellent fit to the data: χ^2 (3) = 4.80, p = .44, RMSEA = .000 (90% CI = .000 - .124), CFI = 1.000, SRMR = .038. There was an insignificant treatment effect at t4 (addition: $d = 0.01$, p = .93; subtraction: $d = -0.08$, p = .66).

Approximate number sense

First and second posttest. For ANS, we achieved measurement invariance, and ANS was represented by two indicators (digit comparison and dot comparison) at each measurement occasion. Allowing for residual covariances between the digit comparison tasks between the three time points resulted in an excellent model fit: χ^2 (3) = 0.77, p = .86, RMSEA = .000 (90% CI = .000 - .083), CFI = 1.000, SRMR = .013. In the second step, we examined the treatment effects by specifying the structural model shown in Figure 4. This model exhibited a very good fit to the data: χ^2 (8) = 9.43, p = .31, RMSEA = .039 (90% CI = .000 - .118), CFI = .993, SRMR = .050. Overall, the treatment effect at t2 was marginally significant (d = 0.35, p = .06); the treatment effect at t3 was insignificant ($d = 0.22$, p = .25).

Follow-up test 6 months after intervention. Like the structural models for time points t1 to t3, the structural model for time points t1, t2, and t4 showed an excellent fit to the data: χ^2(8) = 7.9, p = .44, RMSEA = .000 (90% CI = .000 - .107), CFI = 1.000, SRMR = .048. This model was based on the assumption of metric invariance over time. Again, the intervention effects at t4 faded out ($d = 0.29$, p = .35).

Discussion

Overall, the intervention produced positive benefits (d = 0.20) on early numeracy learning, but these were not significant. There were moderate and significant effects on

word problem solving ($d = 0.41$), but after correcting for multiple significance tests, results at posttest for word problems were no longer significant. In addition, effects on all four outcome measures were reduced and faded out at the second test (after the second intervention phase) and at the follow-up test (6 months after the intervention) compared to the immediate posttest. Fade-out effects indeed took place for all four outcome variables, indicating that the second phase of the intervention did not successfully prevent or ameliorate such effects.

Immediate Intervention Effects and Transfer Effects

Given the efforts to construct and implement an intervention for children who struggle with numeracy skills, our results can be considered somewhat disappointing. There was a significant effect on word problems, but as mentioned, this was no longer significant after controlling for multiple significance tests. However, it should be noted that whether such a procedure should be employed is debatable (e.g., see Gelman et al., 2012; Rothman, 1990). Indeed, it has been argued that these kinds of correction procedures are too conservative and lead to elevated levels of type 2 errors (Rothman, 1990).

As for the reasons behind the rather weak effect found in our study, one is that the power level was based on overly optimistic assumptions of how large the effects would be, and we did not have sufficient power to detect the effects around 0.2 to 0.3 Cohen's *d*. Another reason is that the duration of the intervention in our study was 8 weeks, three times per week. Although this intervention intensity was comparable to many other studies (see Chodura et al., 2015; Dennis et al., 2016), our intervention study comprised a mere 24 sessions compared to 48 sessions in Fuchs et al. (2005), an additional 30 hours on top of typical classroom instruction in Gersten et al. (2015), and 50 sessions in the "ROOTS" program by Clarke et al. (2016). A third reason could be that, since the children did not have particularly severe mathematical learning difficulties, some of the intervention content may have been too easy for them.

Considering the effect on word problems, it should be noted that word problems were not directly trained in the intervention, so this may support theories of knowledge transfer, at least within the same domain (Taatgen, 2013). One reason for this effect could be that improving children's numeracy and arithmetic skills in general will help them solve word problems. It may also be that, in small groups in which the teachers provided explicit instructions, the activity of solving and reasoning about mathematical tasks (and using expressive language to talk about mathematical problems) enhanced the children's quantitative language skills, thereby helping them to solve word problems on their own. However, results for word problems must be interpreted with the caveat that their effects were no longer present after correcting for multiple significance tests.

Second Intervention Phase

As for follow-up versus immediate effects, the effects in the current study faded when the initial weeks of the intervention had ended. In previous studies examining intervention in young at-risk children, only one of the randomized controlled trials has reported results on follow-up effects (Clarke et al., 2016). This trial also showed clear fade-out effects. This is particularly problematic for interventions that build early numeracy skills because most children are likely to eventually acquire at least minimal levels of these skills soon after entering school. Indeed, much of the fade-out effect in early childhood interventions has been attributed to this type of catch-up among the larger population of children (Bailey et al., 2016). Thus, fade-out effects have important implications for teaching. Early interventions do not imply that the children's challenges are solved but that children who experience problems are likely to need interventions regularly so that they do not fall back into a lower developmental trajectory.

As for the reasons why interventions fade out, our study was not designed directly to examine the nature of fade-out effects. Considering the constraining content hypothesis (Bailey et al., 2016), our study attempted to sustain the intervention effect by adding a second intervention phase. However, our study did not incorporate environmental factors such as how teachers could build on the skills the children had learned after the intervention ended or how they were instructed in ordinary classroom settings. We also did not plan how teachers could sustain the effects after the intervention ended. Furthermore, the preexisting differences hypothesis (Bailey et al., 2016) suggests that fade out is due to stable, underlying characteristics in mathematics that cause children to revert to their previous individual trajectories (Bailey et al., 2016). This notion makes it hard to ameliorate children with mathematical learning difficulties, and this may also be related to the intensity of our intervention. To tackle these preexisting differences, the intervention should be maintained for longer if the new trajectories are to be sustained.

Recommendations for Future Studies and Conclusion

In future studies, it will be important to conduct more well-controlled and well-powered studies to increase our knowledge about how difficulties in learning mathematical skills

can be prevented and ameliorated. Moreover, the fade-out effects in this and other studies underline that future studies should be designed with interventions featuring more sessions and over longer periods, as well as interventions that pause for a certain period to discover how more persistent effects might be achieved. For instance, an intervention could be implemented in blocks (with multiple periods of intervention phases) to see if it is possible to prevent fade out. Indeed, one recent language intervention has successfully applied such a procedure (Hagen et al., 2017). Furthermore, it could also be important with active control groups to control for nonspecific effects (i.e., that the intervention group is given more attention than the control group).

Early numeracy skills and early mathematical development can be seen as gatekeeper skills. Children with low performance in early numeracy are at risk of facing learning difficulties in mathematics. From this and other studies, it is clear that mathematical skills can be improved despite high stability in rank order between children. However, it is important to note that improvement will require great effort, and that most studies have inadequate intervention intensity to achieve this, particularly in the long run. Even though our study included a repetition phase, this was not sufficient to gain lasting improvements. Thus, to achieve this, future studies are likely to need a new continuous take on interventions, with only short breaks in between each phase. Ultimately, it seems unlikely that this type of 10- to 20-week intervention will be helpful for most children who struggle.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material is available on the webpage with the online version of the article.

References

Aunio, P., Mononen, R., & Lopez-Pedersen, A. (2016). *Early numeracy test* [Unpublished].

Aunio, P., & Räsänen, P. (2016). Core numerical skills for learning mathematics in children aged five to eight years—A working model for educators. *European Early Childhood Education Research Journal*, *24*(5), 684–704. https://doi.org/10.1080/1350293X.2014.996424

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

Bailey, D. H. (2019). Explanations and implications of diminishing intervention impacts across time. In D. C. Geary, D. B. Berch, & K. M. Koepke (Eds.), *Cognitive foundations for improving mathematical learning* (Vol. 5, pp. 321–346). Academic Press.

Bailey, D. H., Nguyen, T., Jenkins, J. D., Domina, T., Clements, D. H., & Sarama, J. S. (2016). Fadeout in an early mathematics intervention: Constraining content or preexisting differences. *Developmental Psychology*, *52*, 1457–1469. https://psycnet.apa.org/doi/10.1037/dev0000188

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Brigstocke, S., Moll, K., & Hulme, C. (2016). *Test of Basic Arithmetic and Number Skills (TOBANS)*. Oxford University Press.

Candlish, J., Teare, M. D., Dimairo, M., Flight, L., Mandefield, L., & Walters, S. J. (2018). Appropriate statistical methods for analysing partially nested randomised controlled trials with continuous outcomes: A simulation study. *BMC Medical Research Methodology*, *18*(1), Article 105. https://doi.org/10.1186/s12874-018-0559-x

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

Clarke, B., Doabler, C., Smolkowski, K., Nelson, E. K., Fien, H., Baker, S. K., & Kosty, D. (2016). Testing the immediate and long-term efficacy of a tier 2 kindergarten mathematics intervention. *Journal of Research on Educational Effectiveness*, *9*, 607–663. https://doi.org/10.1080/19345747.2015.1116034

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-Based Early Maths Assessment. *Educational Psychology*, *28*, 457–482. https://doi.org/10.1080/01443410701777272

Dennis, M. S., Sharp, E., Chovanes, J., Thomas, A., Burns, R. M., Cister, B., & Park, J. (2016). A meta-analysis of empirical research on teaching students with mathematical learning disabilities. *Learning Disabilities Research & Practice*, *31*(3), 156–168. https://doi.org/10.1111/ldrp.12107

Doabler, C. T., Clarke, B., Kosty, D. B., Baker, S. K., Smolkowski, K., & Fine, H. (2016). Effects of a core kindergarten mathematics curriculum on the mathematics achievement of Spanish-speaking English learners. *School Psychology Review*, *45*(3), 343–361.

Duncan, G. G., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Epskamp, S. (2015). semPlot: Unified visualizations of structural equation models. *Structural Equation Modeling*, *22*(3), 474–483. https://doi.org/10.1080/10705511.2014.937847

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, *97*(3), 493–513. https://doi.org/10.1037/0022-0663.97.3.493

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Schatschneider, C., Hamlett, C. L., . . . Changas, P. (2013). Effects of first-grade number knowledge tutoring with contrasting forms of practice. *Journal of Educational Psychology*, *105*, 58–77. https://doi.org/10.1037/a0030127

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning difficulties: A meta-analysis of instructional components. *Review of Educational Research*, *79*, 1202–1242. https://doi.org/10.3102/0034654309334431

Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children*, *78*, 423–445. https://doi.org/10.1177%2F001440291207800403

Gersten, R., Rolfhus, E., Clarke, B., Decker, L., Wilkins, C., & Dimino, J. (2015). Intervention for first graders with limited number knowledge: Large-scale replication of a randomized controlled trial. *American Educational Research Journal*, *52*, 516–546. https://doi.org/10.3102/0002831214565787

Hagen, Å. M., Melby-Lervåg, M., & Lervåg, A. (2017). Improving language comprehension in preschool children with language difficulties: A cluster randomized trial. *The Journal of Child Psychology and Psychiatry*, *58*(10), 1132–1140. https://doi.org/10.1111/jcpp.12762

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

Jitendra, A. K., Lein, A. E., Im, S.-h., Alghamdi, A. A., Hefte, S. B., & Mouanoutoua, J. (2018). Mathematical interventions for secondary students with learning disabilities and mathematics difficulties: A meta-analysis. *Exceptional Children*, *84*(2), 177–196. https://doi.org/10.1177%2F0014402917737467

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcome. *Developmental Psychology*, *45*, 850–867. https://doi.org/10.1037/a0014939

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

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2020). *semTools: Useful tools for structural equation modeling* (R Package Version 0.5-3). https://CRAN.R-project.org/package=semTools

Khojasteh, J., & Lo, W.-J. (2015). Investigating the sensitivity of goodness-of-fit indices to detect measurement invariance in a bifactor model. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 531–541. https://doi.org/10.1080/10705511.2014.937791

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kroesbergen, E. H., & Van Luit, J. E. H. (2003). Mathematics interventions for children with special educational needs: A meta-analysis. *Remedial and Special Education*, *24*, 97–114. https://doi.org/10.1177%2F07419325030240020501

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of American Statistical Association*, *83*, 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Lopez-Pedersen, A., Mononen, R., Korhonen, J., Aunio, P., & Melby-Lervåg, M. (2021). Validation of an early numeracy screener for first graders. *Scandinavian Journal of Educational Research*, *65*(3), 404–424. https://doi.org/10.1080/00313831.2019.1705901

Marsh, H. W., Hau, K.-T., & Grayson, T. D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 275–340). Lawrence Erlbaum.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge/Taylor & Francis Group.

Moeller, K., Pixner, J., Zuber, J., Kaugmann, L., & Nuerk, H.-C. (2011). Early place-value understanding as a precursor for later arithmetic performance—A longitudinal study on numerical development. *Research in Developmental Disabilities*, *38*, 1837–1851. https://doi.org/10.1016/j.ridd.2011.03.012

Monei, T., & Pedro, A. (2017). A systematic review of interventions for children presenting with dyscalculia in primary schools. *Educational Psychology in Practice*, *33*(3), 277–293. https://doi.org/10.1080/02667363.2017.1289076

Nguyen, T. T., Watts, G., Duncan, D., Clements, J., Sarama, C., Wolfe, C., & Spitler, M. (2016). Which preschool mathematics competencies are most predictive of fifth grade achievement? *Early Childhood Research*, *36*(3), 550–560. https://doi.org/10.1016/j.ecresq.2016.02.003

Purpura, D. J., Baroody, A. J., & Lonigan, C. J. (2013). The transition from informal to formal mathematical knowledge: Mediation by numeral knowledge. *Journal of Educational Psychology*, *105*, 453–464. https://doi.org/10.1037/a0031753

Putnick, D. L., & Bornstein, M. L. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, *1*(1), 43–46.

Schulz, K. F., Altman, D. G., & Moher, D., & the CONSORT group. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. *BMC Medicine*, *8*, Article 18. https://bmcmedicine.biomedcentral.com/articles/10.1186/1741-7015-8-18

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

Sterba, S. K., Preacher, K. J., Forehand, R., Hardcastle, E. J., Cole, D. A., & Compas, B. E. (2014). Structural equation modeling approaches for analyzing partially nested data. *Multivariate Behavioral Research*, *49*(2), 93–118. https://doi.org/10.1080/00273171.2014.882253

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, *102*, 439–471. https://doi.org/10.1037/a0033138

Wang, A. H., Firmender, J. M., Power, J. R., & Byrnes, J. P. (2016). Understanding the program effectiveness of early mathematics interventions for prekindergarten and kindergarten environments: A meta-analytic review. *Early Education and Development*, *27*(5), 692–713. https://doi.org/10.1080/10409289.2016.1116343

Wechsler, D., Kaplan, E., Fein, D., Kramer, J., Morris, R., Delis, D., & Maelender, A. (2003). *Wechsler Intelligence Scale for Children: Fourth edition (WISC-IV)* [Assessment instrument]. Pearson.

Zhang, Z., & Yuan, K.-H. (2018). *Practical statistical power analysis using WebPower and R*. International Society for Data Science and Analytics Press.