©American Psychological Association, 2026. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. The final article is available, upon publication, at: 10.1037/bul0000516

[MANUSCRIPT ACCEPTED FOR PUBLICATION AT PSYCHOLOGICAL BULLETIN]

On the Mechanisms of Intervention Effect Fade-Out: A Meta-Analytic Review of Interventions Targeting At-Risk Students' Achievement

Jens Dietrichson[1], Roopali Bhatnagar[2], Trine Filges[1], and Mikkel Helding Vembye[1]

[1]VIVE - The Danish Center for Social Science Research
[2]Department of Management, Society and Communication, Copenhagen Business School

Author Note

Jens Dietrichson https://orcid.org/0000-0003-4178-7682
Roopali Bhatnagar https://orcid.org/0000-0002-7867-6540
Mikkel Helding Vembye https://orcid.org/000-0001-9071-0724

Part of the data were collected from four previously published systematic reviews (Dietrichson et al., 2017; Dietrichson, Filges, et al., 2020; Dietrichson et al., 2021; Filges et al., 2018). The analyses in this study have not appeared elsewhere. All meta-analytic data, analysis code, and research materials (including our full coding of each record) are available at https://osf.io/j6n5z/. None of the authors has a conflict of interest.

All funding for this study was provided by VIVE. We are very grateful to Drew Bailey, Paul Bingley, Emma Hart, John Protzko, Hans Henrik Sievertsen, Tyler Watts, and seminar participants at Copenhagen Education Network Workshop, Society for Research on Child Development 2025 Biennial Meeting, Styrelsen for Undervisning og Kvalitet, and VIVE for helpful comments and discussions. We thank Julie Kaas Seerup for assisting us with the coding and assessment of studies written in German.

Correspondence concerning this article should be addressed to Jens Dietrichson, Herluf Trolles Gade 11, 1052 Copenhagen K, Denmark. E-mail: jsd@vive.dk.

Abstract

The effects of preschool and school interventions typically fade out over time. To assess the implications of such fade-out for returns to interventions and for skill development theories, it is essential to know the mechanisms underlying fade-out. In this study, we reviewed and categorized potential fade-out mechanisms and developed testable hypotheses about the distribution of intervention effect sizes. We then evaluated the hypotheses using meta-analysis and data from 42 studies and 677 effect sizes examining the effects of school interventions targeting at-risk students on standardized tests in mathematics and reading. Our findings show that mechanisms linked to challenges of measuring and estimating effect sizes explain only a small part of fade-out. Fade-out appears to be a genuine phenomenon reflecting real changes in the skills of the intervention participants. Among skill-based mechanisms, we found evidence against the treatment group forgetting the skills they acquired in the intervention. We found the most support for mechanisms predicting that the treatment group's skills would grow, but more slowly than those of the control group after the intervention. The two most plausible mechanisms explain the slower growth by the treatment group's inability to build on the skills acquired during the intervention, and by negative side effects on parental and school investments for the treatment group. Because these mechanisms have different implications for the returns to interventions and for theories of skill development, future research should aim to distinguish their contributions to fade-out. We conclude by discussing study designs that can achieve this goal.

Keywords: fade-out; school interventions; mechanisms; meta-analysis

Public Significance Statement

Knowledge about fade-out mechanisms is important for deciding how much resources should be allocated to interventions, how educational policymakers should prioritize between existing interventions, and for researchers to develop new and better interventions. This study provides an overview of potential mechanisms of fade-out of intervention effects, and tests hypotheses connected to the mechanisms using a sample of school interventions. Overall, fade-out is a genuine phenomenon, and not mainly due to an inability to properly measure or estimate effects. Our findings also indicate that fade-out is not due to intervention participants forgetting their acquired skills. We find most support for mechanisms explaining fade-out by the treatment group being unable to build on the skills acquired in the intervention, and negative side effects on parental and school investments.

On the Mechanisms of Intervention Effect Fade-Out: A Meta-Analytic Review of Interventions Targeting At-Risk Students' Achievement

Interventions in preschool and school can have transformative effects on skills and academic achievement (e.g., Dietrichson et al., 2021; Duncan & Magnuson, 2013; Nickow et al., 2024). However, the intervention effects are typically larger when measured at the end of the intervention than when measured at later occasions (Abenavoli, 2019; Bailey, Duncan, et al., 2020; Bus & Van IJzendoorn, 1999; Camilli et al., 2010; Dietrichson et al., 2021; Duncan & Magnuson, 2013; Li et al., 2020; Protzko, 2015; Suggate, 2016). While such fade-out of effect sizes between end-of-intervention tests (*post-tests*) and later tests (*follow-up tests*) has been thoroughly documented, the mechanisms behind fade-out are not fully understood (Abenavoli, 2019; Bailey, Duncan, et al., 2020; Bailey et al., 2017; Kang et al., 2019; Pages et al., 2023).

Knowledge about fade-out mechanisms is important for assessing the returns to interventions. Although long-term effects may not be necessary to make interventions worth pursuing (Abenavoli, 2019; Bailey, Duncan, et al., 2020; Watts, Bailey et al., 2019; Whitehurst, 2018), lasting effects on skills greatly improve the benefits-to-cost ratio of interventions (Wolf et al., 2019). Mechanisms explaining apparent fade-out by, for example, changes to sample composition or measurement issues, are compatible with fade-out and lasting effects on treatment group skills. Other mechanisms–like the treatment group forgetting what they learnt in the intervention–are not compatible with lasting effects. Thus, assessments of the returns to interventions depend on knowledge about fade-out mechanisms.

Many theories of skill development imply cumulative advantages, or "Matthew effects" (e.g., Cunha & Heckman, 2007; DiPrete & Eirich, 2006; Duncan et al., 2007; Knudsen et al., 2006; Masten & Cicchetti, 2010; Savi et al., 2019; Share, 1995; Stanovich, 1986; Van Der Maas

et al., 2006). If skills acquired early in life help later skill development or early investments increase the effect of later investments, then children "rich" in skills will become more skilled over time, and inequalities will grow (Cunha & Heckman, 2007). In both reading and mathematics, basic skills (e.g., decoding and single-digit arithmetic) build the foundation for more complex skills (e.g., reading comprehension and multi-digit arithmetic; Hoover & Gough, 1990; Hoover & Tunmer, 2018; Watts et al., 2017). This structure seems conducive for Matthew effects (Share, 1995; Stanovich, 1986; Watts et al., 2017). Success may furthermore be self-reinforcing because it improves motivation, and students who are more motivated subsequently learn more (Chapman et al., 2000; Eklund et al., 2013; Pfost et al., 2014; Toste et al., 2020).

All else equal, Matthew effects imply growing effects of interventions that successfully improve skills. Fade-out of intervention effects may therefore seem to contradict theories predicting Matthew effects. However, without knowledge of why effect sizes fade out, such conclusions are premature. As we discuss below, many fade-out mechanisms are in fact compatible with an underlying skill development process that exhibits Matthew effects and with the complete fade-out of effect sizes. In turn, more knowledge about skill development theories may lead to the development of better interventions with more persistent effects (Bailey, Duncan, et al., 2020).

This study examines mechanisms that may cause fade-out. We make three contributions: First, we survey and categorize potential explanations of fade-out in educational interventions. Second, we show that the mechanisms have testable implications and derive hypotheses regarding the distribution of effect sizes. Third, we test the hypotheses using meta-regressions and a sample of 42 studies that examined the effects of school interventions targeting at-risk students on standardized tests in mathematics and reading.

We proceed as follows: in the next section we survey and categorize fade-out mechanisms and discuss their implications. We then discuss related meta-analyses, present the current study in more detail and derive hypotheses for the distribution of effect sizes. In the Method section, we describe the sample, coding procedures, outcome measures, and statistical frameworks used to test the hypotheses. Thereafter, we present the results. Lastly, we discuss the interpretation of the hypothesis tests and their implications for the fade-out mechanisms, and offer concluding remarks and directions for further research.

Mechanisms of Effect Size Fade-Out

Following Machamer et al. (2000), we defined fade-out mechanisms as entities (e.g., schools, parents, children, and researchers) and activities (e.g., teaching, instructing, learning, and publishing) organized in such a way that they produce fade-out of effect sizes. We begin with a discussion of three mechanism categories, illustrating how they explain fade-out through changes in the skills of treatment and control groups, namely skill decay, compensatory catch-up, and treatment-induced catch-up. Next, we discuss three other mechanism categories that attribute fade-out to difficulties of properly measuring or estimating effects: sample composition, regression to the mean (RTM), and measurement issues.

We discuss the implications of a mechanism as if it were the only explanation of fade-out. However, it is important to note that the mechanisms are often not mutually exclusive. We return to this topic when we discuss the interpretation of our results.

Skill-Based Mechanisms

Mechanisms in the three skill-based categories imply that the growth curve of skills is concave, as indicated by Andersen et al. (2022) and Lipsey et al. (2012), but have different im-

plications for the returns to interventions and are compatible with different theories of skill development. Formal models in the vein of Cunha and Heckman (2007) conceptualize skill development as a product of earlier skills, investments, and time-invariant factors (e.g., parental education). In the language of these models, skill decay and compensatory catch-up explain fade-out by the form of the production function governing skill development. We separated skill decay and compensatory catch-up to contrast mechanisms that predict an absolute decline in the treatment group's post-intervention skills and emphasize forgetting (skill decay) with mechanisms that predict a relative decline compared to the control group, which does not necessarily involve forgetting (compensatory catch-up).

In contrast, treatment-induced catch-up explains fade-out by the intervention affecting the post-intervention investment levels in the treatment or the control group, and, like compensatory catch-up, predicts a relative decline. The investment levels are unaffected by the intervention under skill decay and compensatory catch-up.

Skill Decay

A common explanation of fade-out in the earlier literature is that skills decay (Cascio & Staiger, 2012). Researchers frequently include "delayed post-tests" very shortly after the immediate post-test (e.g., Dyson et al., 2013, 2015, 2020; Jordan et al., 2012; Fuchs et al., 2022; Sood & Jitendra, 2021), use, or express the need for, short booster or practice sessions that remind treatment group students about the intervention content (e.g., Lindström-Sandahl et al., 2024; Martin, 2021; Nelson et al., 2021; Vaughn et al., 2008), and develop practices to help students remember intervention content (see e.g., Vlach, 2019). These are signs that fade-out is often associated with skill decay.

Skills may decay because they are not used (Dinerstein et al., 2022), because the new skills were not practiced enough to be automatized (Hallstedt et al., 2018; Logan, 1997; National Mathematics Advisory Panel, 2008) or not properly consolidated in the long-term memory and therefore forgotten (Kang et al., 2019; Wixted, 2004), or because new memories may inhibit the retrieval of old memories (Anderson & Hulbert, 2021). According to our definition of skill decay, effect sizes fade out because the treatment group forget more of what they learned in the intervention period than they learn between post and follow-up.

In sum, the skill decay mechanism contradicts theories predicting Matthew effects. Moreover, because the mechanism implies that effects on skills are not lasting, it is a worst-case scenario for the returns to interventions.

Compensatory Catch-Up

Compensatory catch-up mechanisms[1] describe a scenario in which effect sizes fade out because the treatment group is unable to build on the skills learned in the intervention, which the control group eventually learns during regular instruction (Bailey, Jenkins, Alvarez-Vargas, 2020; Kang et al., 2019). The literature on preschool and school interventions includes two overarching explanations of control group catch-up (Burchinal et al., 2022). One explanation emphasizes the transfer of skill training and the second, often named the "sustaining environments hypothesis", conjectures that the persistence of intervention effects is contingent on the quality of the subsequent learning environment (Bailey et al., 2017).

Training one set of skills does not necessarily affect other skills (e.g., James, 1890; Thorndike & Woodworth, 1901). Transfer occurs, for instance, when training one skill develops

[1] We have borrowed the name from biology, where compensatory growth is an accelerated growth of an organism after a period of slowed development. In our case, the slowed development should be understood as the control group's development *relative* to the treatment group's development between pre and post-intervention tests.

operators or cognitive routines that are useful for other skills (Gathercole et al., 2019; Taatgen, 2013), and "near" transfer to similar skills is more likely than "far" transfer to dissimilar skills (Barnett & Ceci, 2002). For example, skills that are malleable, fundamental, and would not have developed without intervention (Bailey et al., 2017), and unconstrained skills (Ansari et al., 2020; Li et al., 2020; McCormick et al., 2021) may be more transferable and fade out less. A related explanation may be that the treatment group forgets some of the skills learned in the intervention and then re-learns the skills during regular instruction.

Two versions of the sustaining environments hypothesis are a) that the quality of the subsequent environment is more important for the treatment group than the control group because high quality is necessary to further the skills taught in the intervention, or b) that subsequent instruction may be redundant if it contains elements that are similar to the intervention (Burchinal et al., 2022; Jenkins et al., 2018). The sustaining environments hypothesis is in line with theories emphasizing sensitive periods of development (e.g., Knudsen et al., 2006; Montessori, 1912; Teicher et al., 2016; Vygotsky, 2012), in the sense that the intervention creates a window of opportunity for further development with subsequent fade-out if this opportunity is missed. Predictions of Matthew effects are often conditioned on investments occurring at the right time (e.g., Cunha & Heckman, 2007; Knudsen et al., 2006), and that future interventions or instruction may have to build on the earlier skills for Matthew effects to appear (Pfost et al., 2014).[2]

[2] The two theoretical perspectives emphasize different aspects but are compatible. For example, a sustaining environment may be one that coordinates post-intervention instruction with the intervention content to increase the chance of transfer between earlier and later skills.

Compensatory catch-up is not compatible with theories that predict what one may call "unconditional" Matthew effects. That is, there may be Matthew effects of interventions targeting the skills that transfer, or of interventions that are followed by an environment that sustains the skills acquired, otherwise, effects fade out.

Treatment-Induced Catch-Up

This category of mechanisms suggests that fade-out is caused by side effects of the intervention. If successful interventions imply that the treatment groups receive less post-intervention help or training, or fewer future interventions, or poorer school quality than they would have had they not participated in the intervention, then effect sizes are expected to fade out. Similarly, if the control group children are provided with extra help or higher school quality because they were in the control group, fade-out is also expected.

Such side effects may come about in many ways, but two main channels involve schools and parents. The differences in investments may be caused by schools and parents reacting to the different post-intervention skill levels, or because the intervention directly affects the investments. Side effects caused by schools could be expected because schools often provide extra help to the lowest achieving students (Andersen et al., 2022; Bai et al., 2020; Carr et al., 2024; Fredriksson et al., 2024; Schroeders et al., 2016). In studies that assign treatments within schools, a successful intervention would also provide schools with effective means to improve the control group's skills, that is, the intervention may affect investments directly.

We are not aware of studies that examine whether the types of school interventions we included have side effects on parents. However, studies using natural or randomized experiments have found effects on parental behavior and investments from changes in their child's schooling conditions that may result in treatment-induced catch-up. Examples include that parents 1) offset

school grants by lowering their own education spending (Das et al., 2013), 2) reduce their efforts when their children attend better schools (Pop-Eleches & Urquiola, 2013) or perform better (Fredriksson et al., 2024), 3) decrease their time investments when they learn from school inspections that their school is of high quality (Greaves et al., 2023), and 4) choose lower quality schools when their children participated in a successful preschool intervention (Currie & Thomas, 2000; Gensowski et al., 2020).

Similar changes after the interventions we study would also cause fade-out. However, this pattern is not universal. For example, some preschool interventions changed parental investments in a way that ought to have increased the chances of persistent or growing effects instead of fade-out (Gelber & Isen, 2013; Kitchens et al., 2020; Reynolds & Ou, 2011; Watts et al., 2020). Indeed, increased parental investments may be one reason why some intervention effects persist long into adulthood (Abenavoli, 2019; Garcia & Heckman, 2023; Reynolds & Ou, 2011). Nevertheless, as the earlier examples show, treatment-induced side effects on parental behavior and investments are a possible reason for treatment-induced catch-up.

As we discuss further in the Hypothesis section, the implications of treatment-induced catch-up may differ depending on whether the side effects affect the investments for the treatment or the control group. Henceforth, we call these two versions treatment-induced (t) and (c). Both treatment-induced catch-up mechanisms are compatible with theories that predict (unconditional) Matthew effects. The reason we do not observe Matthew effects is that the treatment induces later investments with countervailing effects. Furthermore, failing to account for treatment-induced catch-up underestimates the returns to intervention (Abenavoli, 2019; Raudenbush, 2001; Raver, 2012).

Statistical and Measurement Mechanisms

The next three categories of mechanisms explain fade-out by an inability to properly measure or estimate effects. These mechanisms are all compatible with Matthew effects.

Sample Composition

Changes to the composition of the treatment and control groups between post and follow-up tests may explain effect size fade-out (Bailey, Duncan, et al., 2020; Schneider & Bradford, 2020). For example, students with more severe academic difficulties may be less willing to participate in follow-up testing or more likely to be absent, change schools, or drop out between the post and follow-up tests (e.g., Mantzicopoulos et al., 1992). A successful intervention would then imply differential attrition: low-achieving students in the control group would leave the sample between post and follow-up at a higher rate than low-achieving treatment group students, which would increase the control group mean score and cause fade-out (as long as the standard deviation is not decreased enough to make up for the smaller mean difference).[3]

Regression to the Mean

Far from all school interventions are evaluated with both post and follow-up tests (e.g., Dietrichson, Filges, et al., 2020; Dietrichson et al., 2021). A potential implication of this selection process is RTM: if a variable for random reasons shows large or small values on one measurement occasion, then later measurements are likely to be closer to the long-run mean of the variable (Barnett et al., 2005; Linden, 2013). For RTM to explain effect size fade-out, we need two assumptions to hold in addition to not being able to observe all effect sizes at follow-up (Bailey, Duncan, et al., 2020):

[3] High-achieving treatment group students leaving the sample at a higher rate than high-achieving control group students would similarly cause fade-out. However, we are not aware of studies documenting this type of differential attrition, and we therefore did not consider it.

Assumption 1 (A1): The probability of conducting a follow-up measurement is higher for larger post-test effect sizes than for smaller post-test effect sizes.

Assumption 2 (A2): Effect sizes are affected by random factors.

Watts, Bailey et al. (2019) argued that there are incentives to produce and follow up interventions with large short-term effects. If A1 and A2 hold, then interventions that due to random factors had larger post-test effect sizes are more likely to receive a follow-up (Bailey et al., 2017; Watts, Bailey et al., 2019). A potential implication of the combination of A1 and A2 is that conditional on the post-test effect size, effect sizes with larger standard errors are more likely to be over-estimated and should fade out more.

Other selection processes may also cause RTM. For example, if researchers, editors or reviewers prefer to find statistically significant effects (henceforth, "significant" refers to statistical significance only, and not to, e.g., the size of estimates), both sample size and effect size matter for the selection of which interventions to follow-up. Thus, the selection process can be more complex than the simple process sketched above, and fade-out would still be observed.

Measurement Issues

Bailey, Duncan, et al. (2020) described three measurement issues that may cause fade-out. First, at least for some tests, the variance tends to increase with student age (Cascio & Staiger, 2012; Wasik & Slavin, 1993). Effect sizes are commonly calculated by dividing an effect estimate with some version of the standard deviation of the mean test score. Increasing the variance from post to follow-up would therefore result in a lower effect size.

Second, changes in estimated group differences may be influenced by the vertical scaling of tests (e.g., Bond & Lang, 2013; Protopapas et al., 2016; Raudenbush, 2001; Student, 2022). Vertical scaling takes students' responses to test items on two tests of a content domain and

places the test scores onto a common scale (Student, 2022). Because the latent construct captured by tests may differ, even for tests that purportedly measure the same skill, fade-out may be caused by post and follow-up tests weighting skills differently (Wan et al., 2021).[4]

Third, if the post-test is researcher-developed and the follow-up test is a standardized test, then, because researcher-developed tests often yield larger effect sizes (Cheung & Slavin, 2016; Scammacca et al., 2015; Wolf & Harbatkin, 2022), fade-out may be the result. As we discuss below, our sample only contained standardized tests and this issue was not pertinent.

An additional measurement-related issue may be the blinding of testers. For example, follow-up tests may more often than post-tests be school or statewide tests that all students take, regardless of whether they participated in an intervention. Testers are unlikely to be aware of students' treatment status when tests are taken independently of interventions, whereas for post-tests, testers are often aware of this status (Dietrichson et al., 2017; Dietrichson, Filges, et al., 2020; Dietrichson et al., 2021). Lack of blinding of treatment status may, if testers want the intervention to succeed, inflate post-test effect sizes and cause fade-out. However, our sample included only one study in which we could ascertain that the blinding of testers differed between the post and follow-up tests (Ryder et al., 2008).[5]

In sum, several potential measurement issues may cause fade-out. However, in our sample, there were only two pertinent issues: 1) the test score variance may increase with student age, and 2) the vertical scaling of tests.

---

[4] Measures with ceiling or floor effects may similarly create fade-out patterns without any change in the skill difference between the treatment and control groups (Pfost et al., 2014). Our risk of bias assessment included an assessment of floor and ceiling effects. No included study was judged to have too high risk of bias in this regard. However, not all studies included enough information to ascertain the degree of ceiling and floor effects.

[5] In Supplemental Material C, we show that effect sizes from studies for which we could ascertain that the blinding was the same for post and follow-up tests also exhibit fade-out. Differential blinding of testers is therefore not a likely explanation of fade-out in our data.

Prior Meta-Analyses and Reviews

A few earlier meta-analyses and reviews have examined fade-out mechanisms.[6] In a pioneering meta-analysis, Protzko (2015) examined fade-out of effect sizes based on IQ measures using a sample of 23 studies of primarily early childhood interventions. Protzko (2015) argued that the fade-out of IQ effect sizes was caused by the treatment group gradually losing their initial advantage over time, rather than by children in the control group improving.

Li et al. (2020) used meta-analysis to examine, among other things, whether the effects of early childhood education interventions were more persistent for achievement or cognitive skills outcomes, for interventions with longer duration, and for interventions implemented earlier in a child's life. There were 18 studies with follow-up measurements in their sample of interventions from the United States (US). The results indicated that achievement and cognitive skills outcomes faded out at a similar rate and that longer interventions and interventions implemented earlier in life had more persistent effects.

Hart et al. (2024) and Rosengarten et al. (2024) meta-analyzed 87 and 43 interventions to examine how fast different types of skills fade out. Hart et al. (2024) found comparable fade-out of cognitive and social-emotional skills during the first year after the intervention. At one to two-year follow-ups, effects on cognitive skills showed slightly less fade-out. Rosengarten et al. (2024) examined constrained and unconstrained skills, and found no significant difference in some models, and more fade-out of unconstrained skills in others.

Bailey, Duncan, et al. (2020) conducted a narrative review of the fade-out and persistence of effects of a large set of primarily educational interventions. They discussed a partly overlap-

[6] We describe evidence from primary studies that tested implications and assumptions of fade-out mechanisms when we discuss our results in the Discussion-section.

ping set of mechanisms as our review, and concluded that 1) although fade-out is the norm, it often co-exists with persistence; 2) fade-out may to some extent be a measurement artifact but is unlikely to be fully explained by measurement issues; and 3) persistence depends on the types of skills targeted by the intervention, institutional constraints and opportunities, and complementarities between interventions and subsequent environmental affordances. An example of subsequent environmental affordances, which our more uniform sample did not include, is gateway mechanisms, such as interventions that help students make the transition into college. These interventions have often shown persistent effects.

Abenavoli (2019) provided a narrative analysis of 53 articles about moderators and mechanisms of fade-out in early childhood education programs implemented in the US. There was relatively consistent evidence that effect sizes were more persistent for children at greater sociodemographic risk (but less consistent for other risk factors such as low baseline skills). There was evidence of early cognitive, academic, social-emotional, and motivational skills, as well as family and school support mediating the longer-term effects.

The Present Study

Our overarching research question for this study was: Which are the most important mechanisms of effect size fade-out in school interventions? Our survey discussed potential mechanisms, including those that might be causing fade-out in our data. Below, we describe implications of these mechanisms by formulating eight hypotheses that can be tested with our data. We have also described assumptions of the RTM and the increasing test variance mechanisms, for which we also provide tests.

We used data from four systematic reviews of school interventions implemented in kindergarten (K) to grade 12 (Dietrichson et al., 2017; Filges et al., 2018; Dietrichson, Filges, et al.,

2020; Dietrichson et al., 2021). Our sample consisted of the subset of studies from these reviews that included a post-test and at least one follow-up outcome plus studies found by forward citation-tracking these studies. Outcomes were standardized tests of reading or mathematics with a sufficiently low risk of bias. All interventions were conducted in an OECD country and included a majority of students that were at risk of academic difficulties.

We chose this data set for a number of reasons. Albeit not a systematic review, the data set was the result of a very large search effort. The combined searches and citation-tracking included 46,475 records screened on title and abstract, 5,797 records screened in full text, and 1,187 included studies. The homogeneity of the included interventions helped hold factors constant that might otherwise have caused confounding. Furthermore, the source reviews followed the guidelines of the Campbell Collaboration (The Methods Group of the Campbell Collaboration, 2019). At least two persons screened each record on title and abstract and in full text, coded each record, and conducted a risk of bias assessment of each outcome.

To the best of our knowledge, no previous meta-analysis has examined similar hypotheses about the mechanisms behind the fade-out of school interventions. Nor have previous meta-analyses tested multiple mechanisms simultaneously. Because some fade-out mechanisms directly concern the distribution of studies and effect sizes (RTM) or are difficult to test in primary studies (sample composition), we believe meta-analysis is an important complement to primary studies on fade-out.

Lastly, while our familiarity with some of the studies was instrumental in developing some of the hypotheses, it also meant that we could not pre-register the hypotheses and analyses. Thus, our analysis is exploratory.

Hypotheses

In this section, we derive a set of hypotheses and discuss what rejecting or failing to reject these hypotheses would imply for the categories of mechanisms relevant to our data. With the exception of increasing variance and vertical scaling of tests, these categories contain more than one mechanism (i.e., more than one example of entities acting in a way that produces fade-out). To ease the reading, we will nevertheless use "mechanism" rather than "categories of mechanisms" throughout. However, it is important to note that the hypotheses presented below provide evidence for or against all mechanisms within a category.

Table 1 summarizes the connections between the mechanisms and the hypotheses. We first present four hypotheses (H1-H4) that we tested with our full data set and elaborate on these connections.

H1 (Sample Composition). *There is no or less fade-out in studies where the sample is identical at the post and follow-up tests.* H1 tests whether changes to the sample composition are an important explanation of effect size fade-out by examining studies where such changes cannot be the reason for fade-out because the samples used at post and follow-up were identical. If we find fade-out with identical samples, then changes to the sample composition are not the only reason for fade-out. If we find less fade-out in these studies than in other studies, then changes to the sample composition may explain some fade-out.

H2 (Follow-Up Duration). *Fade-out is unrelated to the duration between the post and follow-up tests*. The three skill-based mechanisms and increasing test variance are not compatible with H2. All include explanations of fade-out (e.g., forgetting, skill transfer, investments, student heterogeneity) that are more likely to occur as time goes by. In contrast, if RTM fully explains fade-out and unless there is additional selection (e.g., also of follow-up effect sizes), we should

see a pattern where effect sizes drop to their mean at first follow-up and then stay there at further follow-up occasions. Thus, fade-out should be unrelated to the duration between post and follow-up, in line with H2. Changes to the sample composition are compatible with any functional form of the association between effect sizes and the post to follow-up duration, and so is the vertical scaling of tests (scaling problems may compound over time but it is unclear whether that would lead to more or less fade-out).

H3 (Post-Test Effect Size). *Effect size fade-out is unrelated to the post-test effect size*. Both versions of treatment-induced catch-up explain fade-out by parents or schools reacting to an intervention that improves skills in the treatment group. Because a larger effect is more likely to be observed by parents and schools, we expect a stronger reaction to interventions that have larger post-test effects and thus more fade-out in absolute terms.[7] For treatment-induced (t), larger effects also increase the probability that treated students end up at a skill-level where parents or schools reduce their help. All else being equal, effect sizes affected by positive random "shocks" would be larger, more likely to be observed at follow-up, and more likely to mean revert, compared with other effect sizes with similar non-random components. Thus, the treatment-induced catch-up and RTM mechanisms predict a rejection of H3.

If forgetting is proportional to the skill level (as in Campbell & Frey, 1970), then skill decay would also imply more fade-out, in absolute terms, from interventions with larger post-test effect sizes. Similarly, if only a proportion of skills taught in the intervention transfer to the skills measured at follow-up (as in Bailey et al., 2018), then compensatory catch-up also predicts more fade-out of larger post-test effect sizes. That is, assuming a proportional functional form of decay

[7] This prediction does not necessarily hold for a relative fade-out measure. There is also an ambiguous relation between the other two skill-based mechanisms and relative fade-out measures. We discuss this issue further in the Analysis-section and at length in Supplemental Material D.

and transfer, skill decay and compensatory catch-up predict a rejection of H3. The other mechanisms do not have implications for the relation between post-test effect sizes and fade-out.

H4 (Overlapping Test Domains). *The degree of overlap between the domains tested at post and follow-up is unrelated to fade-out*. This hypothesis is related to theories that predict more transfer between tests of similar or overlapping skills (i.e., that near transfer is more likely than far transfer). Compensatory catch-up would be more likely when there is less overlap in this sense. None of the other mechanisms depend on the overlap between tested domains.

We tested the next two sets of hypotheses on subsets of our sample, using different outcome measures for each hypothesis.

H5a (Raw Scores, Identical Test Items). *If identical test items are used at post and follow-up, the raw score of the treatment group is stable*. H5a is compatible with the predictions of compensatory catch-up and treatment-induced catch-up, which imply stable or growing raw scores on tests that are identical (i.e., contain the same test items). Skill decay predicts the opposite: the raw scores will decline. Mechanisms explaining fade-out with statistical or measurement issues do not have implications for H5a.

H5b (Effect Sizes, Identical Test Items). *If the raw scores from identical test items are used to calculate effect sizes at post and follow-up, there is no fade-out*. The vertical scaling of tests should not be an issue when the raw scores from identical test items are used to calculate effect sizes (the scale and standard scores of these tests may, on the other hand, be vulnerable to vertical scaling effects). If we find fade-out also on identical test items, such results would be evidence against vertical scaling as an explanation of fade-out. Skill decay, compensatory catch-up, treatment-induced catch-up, RTM and increased test variance all predict the rejection of H5b.

The next two hypotheses compare the development of the treatment and control group with a group of students who did not participate in an intervention. In our sample, this comparison group was usually a sample used for the norm-referencing of standardized tests. In a few of our studies, this group was instead a group of normally achieving students who were tested to provide an outside reference group but who were not assigned to the treatment or control group. The hypotheses are derived under the assumption that the development of the treatment and control groups and the outside samples would have been parallel, had the intervention not happened.

H6a (Difference Outside Sample, Treatment). *The score difference between the treatment group and the outside sample is stable between the post and follow-up tests*.

H6b (Difference Outside Sample, Control). *The score difference between the control group and the outside sample is stable between the post and follow-up tests*.

Skill decay implies that the treatment group's skills decline after the intervention while the control group and the outside sample are unaffected. That is, skill decay predicts a negative development for the treatment group (against H6a) and a stable control group development (in line with H6b) compared to the outside sample. Treatment-induced (t) makes the same predictions based on the treatment group receiving less post-intervention support because of the intervention. Treatment-induced (c) explains fade-out by the control group receiving extra support because of the intervention, and therefore predicts a stable treatment group score difference (in line with H6a). Because of the extra support, the control group difference to the outside sample should become smaller (i.e., against H6b). Compensatory catch-up implies that the control group develops in their normal pace (in line with H6b), whereas the treatment group develops at a slower pace than other students after the intervention as they are unable to build on their newly learnt skills (against H6a).

Changes to the sample composition imply that low-achieving control group students leave the sample at a higher rate than low-achieving treatment group students, which would lead to an increase in the control group mean score and a rejection of H6b. Because the mechanism only requires a higher attrition rate in the control group, it does not constrain the treatment group attrition and makes no prediction regarding H6a.

RTM and vertical scaling of tests predict that effect sizes fade out because the treatment group's score decreases or the control group's score increases (or both). These two explanations are therefore compatible with rejecting at least one of H6a and b, but not with failing to reject both. Increasing test variance would imply that the same absolute difference between an outside sample and the treatment (or control) group results in a smaller effect size when standardized with a larger standard deviation. That is, the relative score difference of treatment and control groups compared with outside samples should decrease between post and follow-up (i.e., a rejection of both H6a and b).

As mentioned, these tests require that the treatment and control groups, on average across samples, would have developed at the same pace as the outside sample had the intervention not occurred. That is, that the outside samples provide a valid counterfactual to the treatment and control groups' developments. This assumption is similar to the parallel trends-assumption needed to identify effects in difference-in-differences analyses (e.g., Abadie, 2005).

Table 1

Summary of predictions and hypotheses by mechanism

Mechanism	H1	H2	H3	H4	H5a	H5b	H6a	H6b
Skill decay		Longer post–follow-up duration, more fade-out	Larger post-test ES, more fade-out		Decreasing T scores on identical test items	Fade-out on identical test items	Increasing T-outside sample difference	Stable C-outside sample difference
Compensatory catch-up		Longer post–follow-up duration, more fade-out	Larger post-test ES, more fade-out	Less overlap, more fade-out	Non-decreasing T scores on identical test items	Fade-out on identical test items	Increasing T-outside sample difference	Stable C-outside sample difference
Treatment-induced (t)		Longer post–follow-up duration, more fade-out	Larger post-test ES, more fade-out		Non-decreasing T scores on identical test items	Fade-out on identical test items	Increasing T-outside sample difference	Stable C-outside sample difference
Treatment-induced (c)		Longer post–follow-up duration, more fade-out	Larger post-test ES, more fade-out		Non-decreasing T scores on identical test items	Fade-out on identical test items	Stable T-outside sample difference	Decreasing C-outside sample difference
Sample composition	No fade-out in identical samples							Decreasing C-outside sample difference
Regression to the mean		Post–follow-up duration unrelated to fade-out	Larger post-test ES, more fade-out			Fade-out on identical test items	Stable or increasing T-outside sample difference	Stable or decreasing C-outside sample difference
Increasing variance		Longer post–follow-up duration, more fade-out				Fade-out on identical test items	Decreasing T-outside sample difference	Decreasing C-outside sample difference
Vertical scaling of tests						No fade-out on identical test items	Stable or increasing T-outside sample difference	Stable or decreasing C-outside sample difference

Note: Blank cells indicate that the mechanism is compatible with any outcome of the hypothesis test. ES = effect size, T = treatment group, C = control group.

We emphasize that this assumption is strong, especially since we for two reasons cannot use the pre-intervention trends as an indirect check of the plausibility of the assumption. First, several studies select participants who score below a cutoff on a pre-test. This selection ought to produce regression to the mean and the control group would therefore be expected to improve more than the outside sample between the pre and post-test. Second, pre and post-tests are typically conducted at different times of the year, which may make the comparison to a norm-referencing sample problematic if tests for this sample are not conducted with a similar interval. In Supplemental Material G, we probe the sensitivity of our estimates to violations of the assumption using estimates of the trends from the earlier literature.

Method

This section describes the sample of studies, coding procedures, outcomes, the analysis, and transparency and openness. We report how we determined our sample sizes, all data exclusions, all manipulations, and all measures in the study.

Sample of Studies

The starting point for our sample were studies included in four published systematic reviews of school interventions (Dietrichson et al., 2017; Filges et al., 2018; Dietrichson, Filges, et al., 2020; Dietrichson et al., 2021). We use the term "studies" to refer to collections of effect sizes that are unlikely to be independent of one another (e.g., because the same students take more than one test). When the effect sizes were reported in several records (articles, reports etc.), a study is a "cluster" of records.

The four reviews used standardized tests of academic achievement as primary outcomes and included randomized-controlled trials (RCTs), quasi-randomized controlled trials (QRCTs),

and quasi-experimental studies (QES). The control group in these studies received a treatment-as-usual condition. Furthermore, interventions were conducted in an OECD country.

Filges et al. (2018) reviewed class size interventions in grade K-12 in regular private, public, or boarding schools. A total of 8,128 records were screened on title and abstract, and 374 were screened in full text. The review included 127 studies, 10 of which were meta-analyzed.

Dietrichson et al. (2017) included interventions targeting low socioeconomic status students in elementary and middle school (grade K-9) that aimed at improving educational achievement and that could be implemented by schools or local stakeholders. A total of 11,807 records were screened on title and abstract, and 1,137 were screened in full text. The review included 193 studies, 101 of which were meta-analyzed.

Dietrichson, Filges et al. (2020) and Dietrichson et al. (2021) included interventions conducted in schools in grades K-6 or 7-12, respectively, and targeting students with or at-risk of academic difficulties. The search and screening procedures were joint for these two reviews. A total of 24,414 records were screened on title and abstract, and 4,247 were screened in full text. The K-6 review included 607 studies, 205 of which were meta-analyzed. The 7-12 review included 247 studies, 71 of which were meta-analyzed.

In addition to the inclusion criteria that the source reviews had in common (school interventions, study design being RCT, QRCTs, or QES; treatment-as-usual controls; OECD countries), we used the following criteria. A) There had to be results from a post-test and at least one follow-up standardized test of reading or mathematics. B) As in the source reviews, we included effect sizes in the meta-analysis if they had sufficiently low risk of bias at post-test.[8] In contrast

[8] See e.g., Egger et al. (2003), Slavin (1995), and Smithers et al. (2018) for arguments in favor of limiting meta-analyses to studies with sufficiently low risk of bias. All four reviews used the same risk-of-bias tool and excluded effect sizes if they were deemed more likely to mislead than inform the analysis (see supplementary material at

to the source reviews, we only required the post-test to have sufficiently low risk of bias because some reasons for giving a too high risk of bias assessment, such as changes to the sample composition due to attrition, concern precisely the mechanisms we want to examine. C) Lastly, we included studies in which a majority of the participants were at-risk students (as in Dietrichson et al., 2017; Dietrichson, Filges, et al., 2020; Dietrichson et al., 2021).

We found additional studies by forward citation-tracking relevant studies from the reviews (backward citation-tracking was conducted in each review). In total, we included 677 effect sizes from 42 studies. In Supplemental Material A, we describe which records are clustered in which studies.

Table 2 and Table 3 present descriptive statistics of this sample. Almost 60% of the included studies were from the US, the rest were from Denmark, Germany, the Netherlands, New Zealand, Norway, Sweden, Turkey, or the UK. Most studies were RCTs (4 were not) that assigned treatment at the individual student level (5 studies assigned schools or classes), were published in peer-reviewed journals (6 were not) and examined interventions in which the average grade was in elementary school (5 were set in middle school, and 1 in high school).

---

https://osf.io/j6n5z/ for information about the search and screening procedures, the risk-of-bias tool, and the assessment of studies and effect sizes).

Table 2

Descriptive percentages for the included studies

Characteristics	Number of Studies	Effect Sizes	(%)
Context			
Denmark	1	7	2.4
Germany	4	40	9.5
Netherlands	2	10	4.8
New Zealand	1	9	2.4
Norway	1	6	2.4
Sweden	6	144	14.3
Turkey	1	4	2.4
UK	2	10	4.8
USA	24	447	57.1
Design			
QES	4	62	9.5
RCT	38	615	90.5
Outlet			
Gray literature	6	127	14.3
Published	36	550	85.7
Instructional method			
Other method(s)	8	109	19.0
Tutoring	34	568	81.0
Test			
Math	21	176	50.0
Reading	23	495	54.8
Combined tests	1	6	2.4
Timing			
Follow-up (< 12 months)	22	110	52.4
Follow-up (12-17 months)	21	184	50.0
Follow-up (> 17 months)	13	80	31.0
Assignment			
Cluster	5	28	11.9
Individual	37	649	88.1
School			

Characteristics	Number of Studies	Effect Sizes	(%)
Elementary (grade 0-5)	37	649	88.1
Middle (grade 6-8)	5	35	11.9
High (grade 9-12)	1	8	2.4
Sample (across post and follow-up test)			
Identical	13	90	31.0

Note. Percentages calculated at the study-level. In some categories, the percentages sum to more than 100% because the categories are not mutually exclusive (e.g., there may be participating students in both elementary and middle school).

Table 3

Descriptive averages for the included studies

Characteristic	Studies	Effect sizes	Average	SD	Range
Sample					
Raw sample size	42	677	303.0	577.8	23-3238
Effective sample size	42	677	188.0	254.7	23-1232
Intervention group (raw)	42	677	144.0	282.9	12-1658
Control group (raw)	42	677	159.0	308.3	12-1580
% Girls in sample	38	650	47.0	6.2	31-58
Attrition from post to follow-up					
% avg. attrition in total sample	26	435	-11.9	10.1	-42-2
% avg. attrition in treatment	26	435	-11.2	9.8	-43-1
% avg. attrition in control	26	435	-12.4	11.2	-42-6
% avg. difference (control - treatment)	26	435	-1.2	5.9	-15-8
Methodological					
Effect sizes per study	42	677	16.1	19.8	2-96

Note. Averages calculated at the study-level. The number of studies is lower for % Girls in sample, as this information was missing in four studies. Attrition is calculated for studies with non-identical samples and information available about attrition levels (five studies had missing or uncertain information). Note that studies can contain both identical and non-identical samples.

Around 80% of the studies examined a tutoring intervention. Almost 55% of the studies included reading tests, and most studies examined samples that were not exactly the same at post and follow-up (13 studies included at least one outcome estimated on identical samples).[9] There were 22 studies with a follow-up less than 12 months after the end of intervention, and 34 studies included a follow-up conducted 12 months or longer after the end of intervention. Most studies had relatively substantial post to follow-up attrition. However, the treatment-control difference in attrition rates was typically small (on average, 1.2% larger in the control group).

In most of our analyses, we restricted the sample further. We describe the motivation for these restrictions in the section Outcomes.

Coding Procedure

All coding in the original reviews was checked by at least two persons. The additional coding for the current study was checked by at least two of the authors. For the two studies in German, coding was checked by the first author and a German-speaking colleague. All discrepancies were solved by consensus, or a third author decided.

Outcomes

All original effect sizes from the four reviews were standardized mean differences, calculated as Hedges' g (Hedges, 1981):

$$g_{kms} = \left(1 - \frac{3}{4N_{kms} - 9}\right) \times \left(\frac{\vartheta_{kms}}{SD_{kms}}\right)$$

[9] One study reported estimates for both identical and non-identical samples (Powell et al., 2021, 2022). We included the identical samples to increase the statistical power in the test of H1 and the sensitivity analyses that were confined to identical samples.

where N_{kms} is the total sample size, ϑ_{kms} the effect estimate, and SD_{kms} is the unadjusted standard deviation–pooled across the treatment and control group–based on outcome k, at measurement timing m, in study s. A positive g represents a beneficial effect of the intervention.

Fade-Out Measure

For the analysis of H1-H4 and H5b, we defined a quantitative measure that describes the magnitude of fade-out. There are at least four non-trivial issues with such measures.

First, which post and follow-up outcomes should be compared? To be able use the effect size variation both within and between studies, we chose to "anchor" individual follow-up effect sizes by comparing them to the post-test effect size with the closest test content within the same study and treatment. If the same test was used at post and follow-up, we used this test as the anchor. If different tests were used, but the post-test covered the same type of skill (e.g., the same math domain), we used this test as the anchor. If there was more than one post-test of the same skill and one was not obviously closer in content than the others, we used the average across tests. Lastly, if the follow-up test covered a skill domain not represented among the post-tests, we used the study average post-test effect size as the anchor.

Second, measures of fade-out could be absolute or relative; both of which have disadvantages. Absolute measures, such as the difference between the follow-up effect size and post-test effect size, may have a lower bound at zero (Cascio & Staiger, 2012). If there is such a lower bound, then small post-test effect sizes may mechanically exhibit less absolute fade-out.

Despite this problem, there were more serious issues with relative measures in our context. Small absolute fade-out may seem very large with a relative measure if the post-test effect size is small. If the post-test effect size is zero, the measure is not defined. As we show in Supplemental Material D (section Relative Fade-Out Measure), both issues were present in our case.

In particular, using a relative measure created a large number of outliers. A related issue is that a relative measure that divide absolute fade-out with a post or follow-up effect size introduce substantial nonlinearities in the association between effect sizes and the relative fade-out measure. These nonlinearities make the association difficult to capture with linear regressions and the number of studies we have at hand. Lastly, as we show in a simple model in Supplemental Material D, with a relative measure the implications of skill decay and compensatory catch-up for the association between post-test effect sizes and fade-out (i.e., H3) are more dependent on the functional form of skill development models.

For these reasons, we used an absolute measure of fade-out in our primary analyses of H1-H4 and H5b, and provide sensitivity analyses with a relative measure. We defined the absolute measure as:

$$F_{ks} = g_{kfs} - g_{kps}$$

where *f* denotes a follow-up test and *p* denotes the post-test of the anchor to follow-up test *k* in study *s*.

Third, how should the development from post-test to follow-up be interpreted? For positive post-test effect sizes, negative values of F indicate fade-out and positive values growth. However, if the anchor is negative, negative values of F indicate that the follow-up effect size was a less negative, or a positive, value. This development may be interpreted in two ways. Negative post-test effect sizes changing to positive follow-up effect sizes may indicate “sleeper” ef-

fects (i.e., growing effects). However, negative effect sizes that become less negative or turn positive may also indicate fade-out of an initially harmful intervention or RTM of an intervention with null effects. Categorizing such effect sizes as growing seems incorrect.[10]

There are five treatments (about 10%) with negative average post-test effect sizes in our sample. As our research question concerned the fade-out of initially successful interventions, we excluded these treatments from our primary analysis. Excluding these treatments and effect sizes meant that two whole studies were excluded. To examine if our results were sensitive to this issue and to zero being a lower bound, we report results from sensitivity analyses using the full sample and samples that we restricted to treatments with progressively larger average post-test effect sizes.

Fourth, how should the variance of the measures be calculated? This issue is pertinent in relation to the question of how to weight studies in the meta-analysis. Typically, meta-analysts use weights based on estimates of the inverse-variance to give more precise estimates more weight (Borenstein et al., 2009). In a random-effects model, the meta-analyst inputs the inverse variance of the effect size and then estimates the within- and between-study variance. For three reasons, we used the inverse sample size of the treatment and the control groups at follow-up as an approximation of the sampling variance of the effect sizes in our models.

First, inverse-variance weights risk introducing bias when the true effect is not zero because they depend on the estimated effect size in question (Hamman et al. 2018; Lin & Aloe, 2021). Second, all included studies lack information about the correlation between the post and follow-up tests, which is needed to calculate the variance of the fade-out measure (What Works

[10] A relative fade-out measure (e.g., g_{kfs}/g_{kps}) implies additional complications as the measure turns positive if both the numerator and the denominator are negative, and negative if either of the numerator and denominator is negative. We discuss this and other issues with relative measures more in depth in the Supplemental Material D.

Clearinghouse, 2021). Thus, we would need to use an estimate, which would introduce uncertainty. Third, using the sample size allowed us to use a more consistent set of weights across outcome measures and reducing the number of differing factors between specifications.

Effect Sizes Based on Raw Scores From Identical Post and Follow-up Tests

To test H5a (Raw scores, identical test items), we calculated an effect size measure, ES^{I}_{tks}, which used the raw scores from identical test items to measure progress or decline between post and follow-up tests for the treatment group (indicated by *t* below):

$$ES^{I}_{tks} = \left(1 - \frac{3}{4 \times n_{tkfs} - 9}\right) \times \left(\frac{\overline{X}_{tkfs} - \overline{X}_{tkps}}{SD_{tkfs}}\right)$$

ES^{I}_{tks} included a similar small sample correction as g, measured the change in average scores from post to follow-up test for the treatment group (the $\overline{X}_{tkfs} - \overline{X}_{tkps}$ term) and standardized this change with the follow-up standard deviation. A positive ES^{I}_{tks} means that the treatment group average score was higher at follow-up than at the post-test. We computed $\text{Var}(ES^{I}_{tks}) = 1/n_{tkfs}$, where n_{tkfs} is the treatment sample size of the given follow-up measure.

To test H5b (Effect sizes, identical test items), we calculated the absolute fade-out measure using the same set of identical test items. We based the identification of the tests that contained the exact same items at post and follow-up on the descriptions in the included records. For the studies included in the primary analysis, we also asked the corresponding author whether our interpretation was correct and obtained a confirmation, or had access to the original test item data and materials, from 15 out of 16. We discuss potential complications arising from start and stop-rules, and provide additional sensitivity analyses in Supplemental Material B.

Effect Sizes Based on Comparisons with Outside Samples

For the analyses comparing the development of the treatment and control group to outside samples, we calculated an analogue to g:

$$ES_{jkms}^{norm} = \left(1 - \frac{3}{4 \times n_{jkms} - 9}\right) \times \left(\frac{\overline{X}_{jkms} - \overline{X}_{km}^{norm}}{SD_{km}^{norm}}\right)$$

where $\overline{X}_{jkms}$ is the mean in group *j* (treatment or control) for outcome *k* at measurement timing *m* in study *s*, $\overline{X}_{km}^{norm}$ is the mean and SD_{km}^{norm} is the standard deviation in the outside sample. A positive ES_{jkms}^{norm} indicated that the treatment or control group average was higher than the outside sample average. Some outside samples were study-specific and consisted of average and high-achieving students from the same sites as the treatment and control group, who did not receive an intervention. Most outside sample scores were standard or scale scores from norm-referenced tests. That is, if a study reported results on a norm-referenced test, we used the mean and standard deviation in the norming sample of this test. Some scores were expressed in the studies as percentile scores. We converted percentile scores to scores with mean 100 and standard deviation 15 by using a psychometric conversion table. We computed $\text{Var}(ES_{jkms}^{norm}) = 1/n_{jkms}$, where n_{jkms} is the sample size linked to $\overline{X}_{jkms}$.

Analysis

We used meta-analytic methods to test the hypotheses. We first describe a few commonalities among these methods, and then describe the specifications we used to test each hypothesis.

Meta-Analysis Methods

Our sample contained different types of interventions, participants, and outcomes, and we expected heterogeneity both within and between studies. We therefore used random-effects models in all our analyses.

Effect sizes and treatments are nested in studies in our data set. Because fade-out may vary both between effect sizes and treatments (i.e., within studies) and between studies, we included all effect sizes in the analysis. However, effect sizes from the same study are unlikely to be independent. The most common reason for dependencies in our sample was that the same students were tested using more than one test–a "correlated" structure (Hedges et al., 2010). Moreover, one study had "hierarchical" structure–non-overlapping samples that were tested on one test–and thirteen studies mixed a correlated structure and a hierarchical structure, that is, they included multiple non-overlapping samples that were tested on multiple tests (Pustejovsky & Tipton, 2022; Van den Noortgate et al., 2013, 2015).

Like most meta-analytic datasets, our sample lacked information about the true within-sample correlation between outcomes. In our primary analysis, we used the correlated-hierarchical effects (CHE-RVE) model developed by Pustejovsky and Tipton (2022). The CHE-RVE model considers the multi-level structure of the data with effect sizes nested in studies, while simultaneously accounting for dependencies from both correlated and hierarchical effects. The model uses robust variance estimation (RVE) to guard against misspecification of the assumed working model (Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015). Models using RVE typically have the best performance in simulations with dependent effect size data containing multiple dependence structures (Fernandez-Castilla et al., 2021; Vembye et al., 2023).

The CHE-RVE model requires a baseline value for the correlation between pairs of effect sizes from the same study (ρ). As Pustejovsky and Tipton (2022), we used $\rho = 0.6$ in our primary analysis. We tested the sensitivity to using $\rho \in \{0, 0.2, 0.4, 0.8\}$. For inference, we used the small sample adjusted standard errors recommended by Tipton (2015) and Tipton and Pustejovsky (2015; specifically, the CR2 adjustment). We reported 95% confidence intervals (CI) based on the small-sample adjusted standard errors and heterogeneity statistics in the form of the estimated within-study standard deviation ($\hat{\omega}$), the between-study standard deviation ($\hat{\tau}$), and the total standard deviation $\left(\hat{\sigma} = \sqrt{\hat{\omega}^2 + \hat{\tau}^2}\right)$.

Clustered Assignment of Treatment

We adjusted outcomes from treatments that were assigned by cluster (schools or classes) by calculating a design effect, which we used to adjust the sample size (Higgins et al., 2022). We assumed equal average cluster sizes in the treatment and control groups when studies did not report this information, and an intra-cluster coefficient (ICC) of 0.1. We assumed a common ICC because studies with clustered assignment lacked this information, which made it difficult to adjust effect sizes individually by study.

Hypothesis Testing

The first four hypotheses concern the absolute fade-out measure and were tested in the same meta-regression framework.

H1-H4. We used the absolute fade-out measure F as the outcome variable and meta-regressions to test hypotheses 1 to 4. We first show results from regressions that include only explanatory variables related to each hypothesis and then include variables related to all four hypotheses simultaneously. That is, we ran regressions of the following kind:

$$F_{ks} = \alpha + \beta X_{ks} + \mu_s + \upsilon_{ks} + \varepsilon_{ks}$$

where F_{ks} is the measure of fade-out of effect size k in study s, α is an intercept, β is a (vector of) parameter(s) to be estimated, X_{ks} is a (vector of) variable(s) relating to the hypothesis in question, μ_s is the study-level random effect (with $\mathbb{E}(\mu_s) = 0$, $\text{Var}(\mu_s) = \tau^2$), υ_{ks} is the outcome-level random effect (with $\mathbb{E}(\upsilon_{ks}) = 0$, $\text{Var}(\upsilon_{ks}) = \omega^2$), and ε_{ks} is the sampling error (with $\mathbb{E}(\varepsilon_{ks}) = 0$, $\text{Var}(\varepsilon_{ks}) = e_{ks}^2$ and $\text{Cov}(\varepsilon_{ks}, \varepsilon_{ls}) = \rho e_{ks}^2$). We assumed that e_{ks}^2 is equal to $1/n_{ks}^T + 1/n_{ks}^C$ (where n_{ks}^T and n_{ks}^C are the sample size in the treatment and control group) and used this together with the CHE model-estimated values of $\hat{\omega}^2$ and $\hat{\tau}^2$ to weight the effect sizes. The exact study weights used in our models can be found in Equation 6 in Vembye et al. (2023).

Regarding H1 (Sample composition), we included an indicator for the sample being exactly the same at the post and follow-up test. We contrasted these identical samples to samples that were either not identical or where the information about the samples was unclear (five studies). A positive and significant estimate $\hat{\beta}$ in the regression above would imply that there was less, and potentially no, fade-out when the sample was identical. We also ran an intercept-only regression with only effect sizes from identical samples. If the intercept in this regression was significantly smaller than zero, we rejected H1.

To test H2 (Follow-up duration), we included two indicators for the duration between the post and follow-up test, one indicating a duration between 12 and 17 months, and one indicating a duration longer than 17 months. The intercept represented duration between 1 and 11 months. We used this specification as at least two indicators are needed to capture nonlinear associations. The duration distribution has a very large spike at 12 months (see further discussion in section Effect Sizes and Fade-Out). This feature prevented us from using more indicators. If one or both of the duration variables were significant, we rejected H2. To examine whether there might be

selection on follow-up tests, we examined associations in the subset of studies with more than one follow-up and show funnel plots based on the method of Pustejovsky and Rodgers (2019).

To test H3 (Post-test effect size), we included the post-test effect size as an explanatory variable. If $\hat{\beta}$ was significantly smaller than zero, larger post-test effect sizes exhibited more fade-out, and we rejected H3.

We tested H4 (Overlapping test domains) by including an indicator for studies using identical tests at post and follow-up. We contrasted identical and non-identical tests because a very large proportion of post and follow-up effect sizes in our sample overlapped in the sense that they tested the same type of skills. If $\hat{\beta}$ was significantly larger than zero, then fade-out was associated with the overlap between testing domains, and we rejected H4.

A potential issue was that identical tests may be more likely to test constrained skills, which may be less transferable than unconstrained skills. There were relatively few clear-cut cases of tests of constrained skills among the identical tests, but we also show results using an indicator of identical tests of unconstrained skills. In line with the definitions of Ansari et al. (2020) and McCormick et al. (2021), we classified tests of calculation, counting, addition, subtraction, number sense, and letter naming/knowledge as tests of constrained skills.

H5 (Identical test items) and H6 (Outside sample) concerned subsets of our data. We tested these two sets of hypotheses in separate frameworks.

H5a and b. To test H5a (Raw scores, identical test items) and H5b (Effect sizes, identical test items), we used the subset of tests that contained the exact same items at post and follow-up and that reported the raw scores. Note that this subset was not equal to the subset with the value one on the indicator Identical test, which we used to test H4. Effect sizes based on identical tests but not reported as raw scores were excluded from the analysis of H5b.

To test H5a, we ran a meta-regression with ES^I as the outcome variable and included only an intercept. If the resulting weighted average effect size was significantly different from zero, we rejected H5a. If the effect size was significantly negative, we interpreted this as evidence in favor of skill decay. This test examined whether treated students forget more than they learn in the period between the post-test and follow-up. Because the duration between post and follow-up was relatively long in some studies, it was possible that students had time to both forget and re-learn intervention content. To gauge whether this scenario was plausible, we also present results from a version of the test where we added studies sequentially, starting with the shortest follow-up duration.

To test H5b, we ran an intercept-only meta-regression with the absolute fade-out measure as the outcome variable. The vertical scaling of tests should not be an issue when the raw scores from identical test items are used to calculate effect sizes and we interpreted a negative and significant coefficient on the intercept as a rejection of H5b.

H6a and b. To test H6a (Difference outside sample, treatment) and H6b (Difference outside sample, control), we used samples of students outside of the intervention as a reference group and compared the development of the treatment groups' (H6a) and control groups' (H6b) scores to these groups. We calculated ES^{norm} at post and follow-up for both the treatment and the control group, and used this variable as the outcome in the following meta-regression:

$$ES_{ks}^{norm} = \alpha + \beta FU + \mu_s + \upsilon_{ks} + \varepsilon_{ks}$$

where FU is an indicator for follow-up tests, and all other terms are defined as before. We ran separate meta-regressions for the treatment and the control groups and rejected H6a and H6b if $\hat{\beta}$ was significantly different from zero in these regressions.

The tests of H6a and H6b rested on the assumption that the treatment and control groups would, on average, have developed in the same pace (in parallel) as the outside sample, had there not been an intervention. Our outside samples were either samples used to norm-reference standardized tests, which contain low-achieving students, or samples of average and high-achieving students, which do not. Because the included studies primarily included low-achieving students, the assumption may be more likely to hold for the norm-referenced samples, and we report a sensitivity analysis using only these samples.

A further threat to the assumption is that there may be a mismatch between the time of the year for the norm-referencing and the post and follow-up tests. For example, if the norm-referenced scores are based on tests conducted throughout the year while the post-test is conducted in December and the follow-up in May, then comparing both post and follow-up to the same norm-referenced scores will make it seem like treated and control students improve in relation to the norm-referencing sample. However, some norm-referenced tests include several comparison points per year and would not have this problem. Because we could not ascertain for which tests this was the case, we provide results excluding effect sizes based on norm-referenced tests and where post and follow-up testing was conducted at different times of the year.

Summary of the Connection Between Mechanisms and Hypotheses. Table 4 summarizes the connection between the mechanisms and the hypotheses. For each mechanism and hypothesis, the table displays the predicted sign of the β. If the prediction is in line with the hypothesis or if a mechanism is compatible with all outcomes of a test, the table displays a 0 or a blank cell, respectively. One feature that the table does not display is that the predictions of RTM and vertical scaling of tests make joint predictions on H6a (Difference outside sample, treatment)

and b (Difference outside sample, control). As discussed in the Hypotheses-section, they are not compatible with failing to reject both H6a and b.

Table 4

Summary of mechanism predictions and hypotheses

Mechanism	H1	H2	H3	H4	H5a	H5b	H6a	H6b
Skill decay		-	-		-	-	-	0
Compensatory catch-up		-	-	+	0, +	-	-	0
Treatment-induced (t)		-	-		0, +	-	-	0
Treatment-induced (c)		-	-		0, +	-	0	+
Sample composition	0							+
Regression to the mean		0	-			-	0, -	0, +
Increasing variance		-				-	+	+
Vertical scaling of tests						0	0, -	0, +

Note. The +, -, and 0 signs indicate the predicted sign of the β according to the mechanisms. Blank cells indicate that the mechanism is compatible with any outcome of the hypothesis test. The hypotheses are: H1: There is no or less fade-out in studies where the sample is identical at the post and follow-up tests. H2: Fade-out is unrelated to the duration between the post and follow-up tests. H3: Effect size fade-out is unrelated to the post-test effect size. H4: The degree of overlap between the domains tested at post and follow-up is unrelated to fade-out. H5a: If identical test items are used at post and follow-up, the raw score of the treatment group is stable. H5b: If the raw scores from identical test items are used to calculate effect sizes at post and follow-up, there is no fade-out. H6a: The score difference between the treatment group and the outside sample is stable between the post and follow-up tests. H6b: The score difference between the control group and the outside sample is stable between the post and follow-up tests.

In two respects, our tests differed from how hypothesis testing in primary studies are often conducted and interpreted. First, our hypotheses are a convenient way of summarizing the implications of the mechanisms, but the formulations may imply that one mechanism predicts a negative coefficient, another a null coefficient, and a third a positive coefficient. It will therefore be necessary for us to consider not only the statistical significance but also the sign and the size of coefficients, as well as the width of confidence intervals. For example, when a mechanism predicts a null result, a narrow confidence interval and a close-to-zero coefficient might provide relatively strong evidence in favor of that mechanism, whereas a relatively large coefficient with a wide confidence interval that includes zero would not be as strong–even if both tests fail to reject the hypothesis.

Second, for most hypotheses, there was more than one possible way of formulating the test and we provide sensitivity analyses alongside the main specifications and in the section Sensitivity Analyses. Our assessment of the mechanisms therefore rested on several hypothesis tests, and we considered the robustness of the results across specifications.

Transparency and Openness

We adhered to the MARS guidelines for meta-analytic reporting (Appelbaum et al., 2018). All meta-analytic data, analysis code, and research materials, including our coding of the studies, descriptions of the search and screening process, and the risk of bias tools are available at https://osf.io/j6n5z/ (Dietrichson et al., 2025). Data were analyzed using R (R Core Team, 2021) for all our statistical analyses. We used the R packages *metafor* (Viechtbauer, 2022), *clubSandwich* (Pustejovsky, 2022), and *robumeta* (Fisher et al., 2017) to conduct the meta-analysis. This review project was not preregistered. Generative artificial intelligence was not used in this research.

Results

This section presents estimates and distributions of the post and follow-up effect sizes and the fade-out measure, and then the outcomes of hypothesis tests. Thereafter, we show results of the tests of assumptions of the RTM and increasing test variance mechanisms, and discuss the results of the sensitivity analysis.

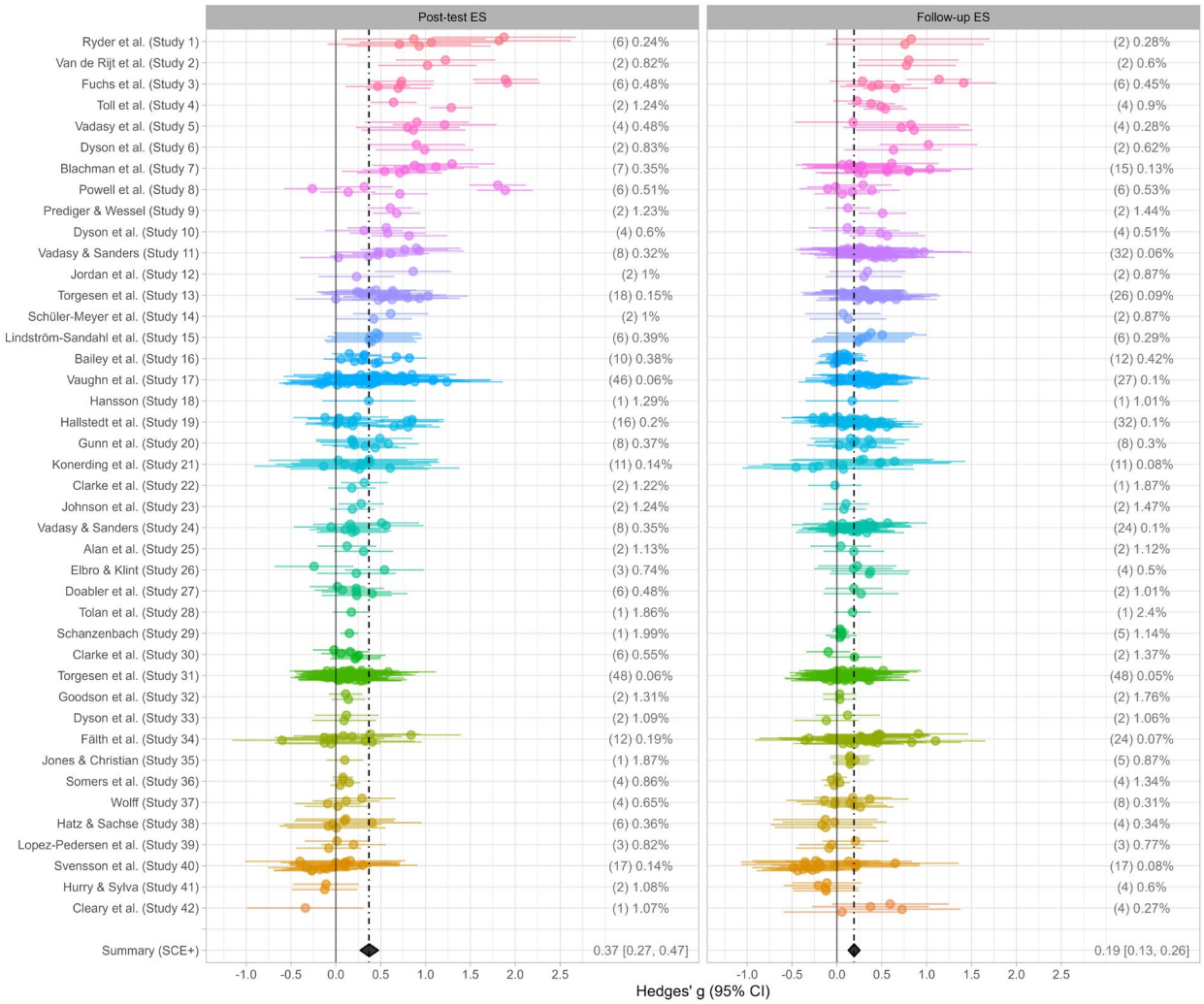
Effect Sizes and Fade-Out

Figure 1 shows a forest plot of the effect sizes by study for post (left panel) and follow-up tests (right panel) using the full sample. We estimated the weighted average effect sizes at post and follow-up reported in the figure using a three-level subgroup correlated effects model (SCE+). This model is similar to the CHE-RVE model but allows for estimations of subgroups within the same model, which therefore can be statistically compared (Pustejovsky & Tipton, 2022). Note that we have omitted two outliers from Figure 1 and Figure 2 to make them more legible. The outliers were included in our other estimations, including the estimates of the weighted average effect sizes presented in text in this section. As can be seen by comparing the coefficients and confidence intervals in Figure 1 to the ones reported below, the influence of outliers on the estimates was very small.

The weighted average, including outliers, was 0.38 at post-test (95% CI = [0.28, 0.49]), and 0.19 at follow-up (95% CI = [0.13, 0.26]). The difference between the weighted averages was significant in the SCE+ model ($p < 0.001$), and the results indicated substantial fade-out on average. The median duration between post and follow-up was 12.0 months, and the weighted average (also estimated with the CHE-RVE model) was 15.2 months.

Figure 1

Forest plot of effect sizes at post-test and follow-up



Note: The figure shows a forest plot of the effect sizes by study for post (left panel) and follow-up tests (right panel). We estimated the weighted average effect sizes at post and follow-up, shown by the diamond at the bottom of each panel, using a three-level subgroup correlated effects model (SCE+). Studies are arranged by their average post-test effect size in both panels (largest at the top). The numbers in parenthesis on the right-hand side of each panel show the number of effect sizes in each study. The percentage indicate each effect size's weight in the

analysis. The weight for each study is therefore the number of effect sizes multiplied by the effect size weight. The study weights sum to 100%.

The figure also indicates some heterogeneity, both within and between studies, which was corroborated by the estimates of the within-study, between-study, and total standard deviation (post: $\hat{\omega} = 0.25$, $\hat{\tau} = 0.31$, $\hat{\sigma} = 0.40$; follow-up: $\hat{\omega} = 0.11$, $\hat{\tau} = 0.14$, $\hat{\sigma} = 0.18$).

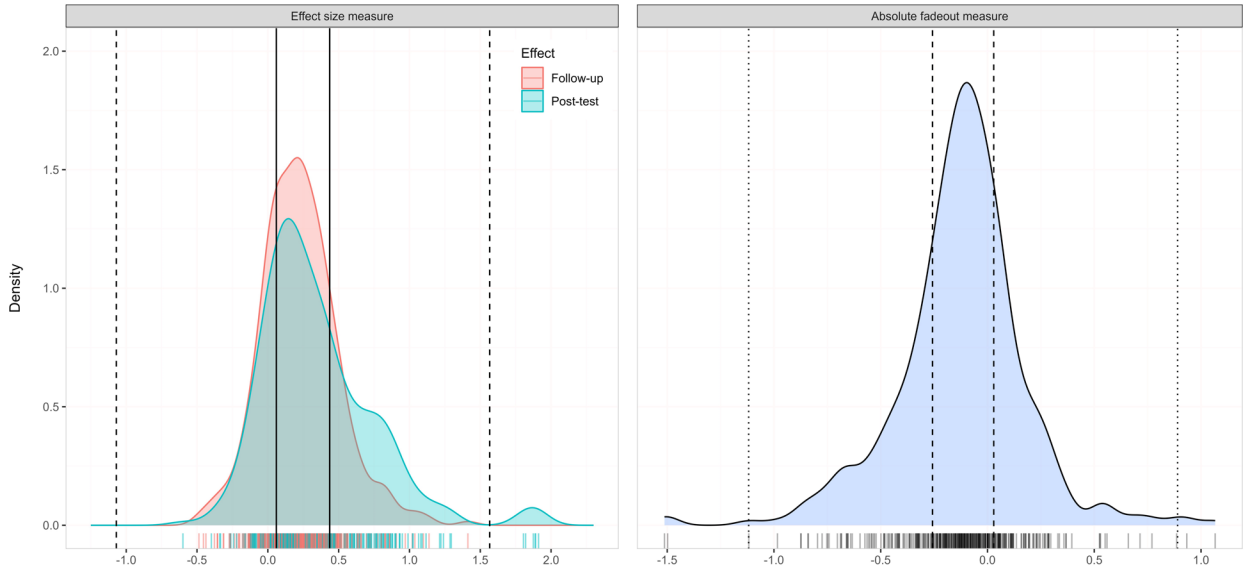
The heterogeneity of both effect sizes and fade-out is clearer in Figure 2. The follow-up effect size distribution is more compressed than the post-test distribution in the left panel of Figure 2. There are fewer large positive effect sizes in the follow-up distribution than in the post-test distribution, whereas the distributions are almost completely overlapping regarding negative effect sizes. The fade-out is concentrated in the range from zero to -0.25 (the 25th percentile is -0.26, and 75th is 0.03). However, there are many observations of substantially larger fade-out and also some observations that indicate persistent or growing effects. The weighted average of the fade-out distribution, estimated with the CHE-RVE model, was -0.15 (95% CI = [-0.21, -0.09]).

As we explained in the Outcomes section, our analyses of H1-H4 used a sample restricted to treatments with on average positive effect sizes. This restriction excluded two studies completely and resulted in almost exactly the same weighted average fade-out of -0.15 (95% CI = [-0.22, -0.09]). The median and weighted average duration between post and follow-up tests were also very close (12 and 16.3 months).[11]

[11] Note that these estimations were based on the post-test effect sizes that were anchors for at least one follow-up effect size, whereas the estimates reported earlier in this section, in text and in Figure 1, included all post-test effect sizes. This difference is a likely explanation of the slight discrepancy between the weighted average fade-out (-0.15) and the difference between the estimated weighted average effect sizes of 0.38 and 0.19 at post and follow-up, respectively.

Figure 2

Distribution of effect sizes and fade-out measure

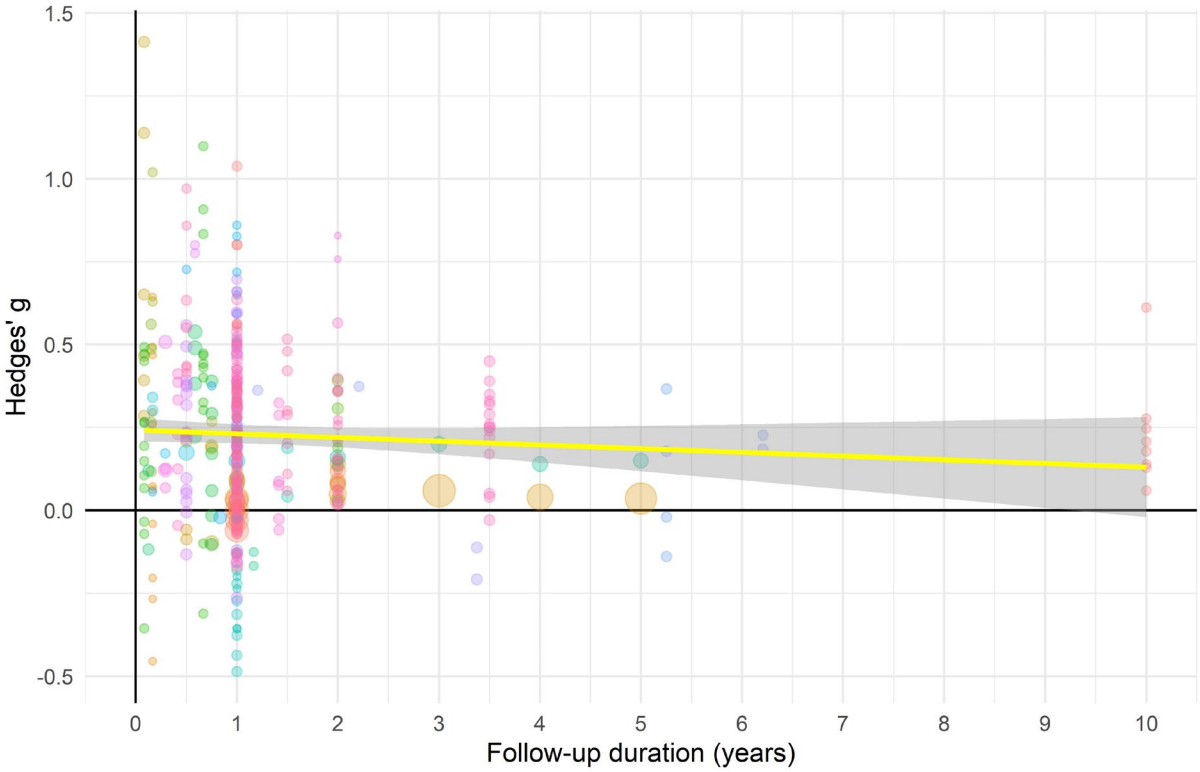


Note. The left panel shows the distribution of effect sizes at post and follow-up. The right panel shows the distribution of the fade-out measure (F_{ks}). The solid lines show the interquartile range and the dashed lines show three times the interquartile range above the first and third quartiles, respectively.

Figure 3 displays how the follow-up effect sizes (y-axis) are distributed over the time elapsed since the post-test (the duration, measured in years on the x-axis). The fitted linear regression slope in Figure 3 appeared quite modest, which implied that most of the fade-out happened early after the post-test. Other noteworthy aspects of Figure 3 were the concentration of follow-up effect sizes in the first year, and especially, exactly at the one-year mark. The figure also highlighted that there was heterogeneity within studies.

Figure 3

Effect sizes and years elapsed since post-test



Note. The displays how the follow-up effect sizes (y-axis) are distributed over the time elapsed since the post-test (the duration, measured in years on the x-axis). The color of the bubbles indicates the study an effect size belongs to, and the size of the bubble indicates the weight of the effect size. The yellow line shows a fitted linear regression slope.

Hypotheses 1-4

Table 5 displays the results of the tests of H1-H4 on the sample including only treatments with, on average, positive post-test effect sizes. Columns 1-4 contain results from specifications in which we included the one or two moderators corresponding to a certain hypothesis, whereas the specification in column 5 included all moderators. Across specifications, the within-study

variation and, in particular, the between-study variation were low. In the specifications that included the post-test effect size, the between-study heterogeneity was zero. That is, there was little heterogeneity that additional study-level moderators could explain.[12]

H1 (Sample composition) stated that there is no or less fade-out when the sample composition did not change between the post and follow-up test. The coefficient on the indicator for identical samples changes sign between columns 1 (negative) and 5 (positive) in Table 5, but both coefficients are insignificant. The weighted average effect size in identical samples was negative and significant ($\hat{\beta}$ = -0.25, 95% CI = [-0.44, -0.07], based on 54 effect sizes and 12 studies). Thus, there was fade-out in identical samples and the magnitude of fade-out was not significantly different in identical and non-identical samples, and we rejected H1.

H2 (Follow-up duration) stated that the fade-out of effect sizes is unrelated to the duration between the post and follow-up test. Both coefficients on the duration indicators are negative and insignificant in column 2 and 5 in Table 5.[13] When we restricted the sample to studies with more than one follow-up (168 effect sizes and 9 studies), the coefficients were more negative and significant in some specifications (around -0.10 and -0.20 in all specifications, and significant when included alone or with the post-test effect size but not with all other moderators). In sum, although we failed to robustly reject the hypothesis that fade-out is unrelated to the duration between post and follow-up tests, the coefficients were large in relation to overall fade-out. There

[12] The estimated between-study heterogeneity is zero for values of ρ down to 0.4 and estimates are small also for 0.2 (see Supplemental Material D, Figure S5). That is, the result holds for reasonable assumptions on ρ.
[13] Joint tests of both coefficients being equal to zero also do not reject the null hypothesis (p = 0.566 in column 2, and p = 0.159 in column 5). Including one combined indicator for a duration of 12 months or longer also yielded a negative and insignificant estimate in the column 5-specification ($\hat{\beta}_{\geq 12}$ = -0.09, 95% CI = [-0.20, 0.01]). Including a continuous (and mean-centered) duration-variable yielded a negative and significant estimate ($\hat{\beta}_{cont}$ = -0.0025, 95% CI = [-0.0046, -0.0004]).

may also have been additional selection of which studies to follow up, which would bias the primary analysis estimates toward zero. In Supplemental Material E we show funnel plots, which are consistent with such selection.

H3 (Post-test effect size) stated that the fade-out of effect sizes is unrelated to the post-test effect size. The coefficients on the post-test effect size variable are negative, large, significant, and of similar size in both columns 3 and 5. Thus, we rejected H3.

H4 (Overlapping test domains) stated that the degree of overlap between the domains tested at post and follow-up is unrelated to fade-out. The coefficients on the indicator for identical tests are positive but small and not significant in columns 4 and 5. Using an indicator of identical and unconstrained tests in the column 5 specification yielded a nearly identical estimate ($\hat{\beta}$ = 0.03, 95% CI = [-0.06, 0.12]). Thus, we failed to reject H4.

Table 5

Tests of hypotheses 1 - 4

	(1) H1		(2) H2		(3) H3		(4) H4		(5) H1-H4	
Moderator	Coef.	df	Coef.	df	Coef.	df	Coef.	Df		df
Intercept	-0.16 [-0.24, -0.08]	26.8	-0.14 [-0.26, 0.03]	16.6	-0.21 [-0.26, -0.16]	18.8	-0.17 [-0.25, -0.08]	19.6	-0.16 [-0.26, -0.06]	11.7
Sample not identical	-0.03 [-0.29, 0.24]	3.5							0.06 [-0.12, 0.24]	8.5
Duration (12-17 months)			-0.01 [-0.15, 0.13]	6.1					-0.08 [-0.18, 0.02]	4.2
Duration (>17 months)			-0.08 [-0.26, 0.11]	8.8					-0.13 [-0.27, 0.01]	6.3
Post-test ES					-0.65 [-0.81, -0.49]	7.5			-0.67 [-0.82, -0.52]	7.2
Identical test							0.0001 [-0.19, 0.19]	9.9	0.02 [-0.11, 0.15]	5.3
Effect sizes	341		341		341		341		341	
Studies	40		40		40		40		40	
Between-study SD	0.04		0.09		0		0.06		0	
Within-study SD	0.21		0.21		0.09		0.21		0.08	
Total SD	0.22		0.23		0.09		0.22		0.08	

Note. The outcome variable in all specifications is the absolute fade-out measure. The sample consists of the treatments with positive average post-test effect sizes. The df column contains the small sample adjusted Satterthwaite degrees of freedom. The terms in brackets below the coefficient estimates denotes the 95% confidence interval. The median and weighted average duration between post and follow-up tests were 12 and 16.3 months, respectively, in all specifications.

Hypotheses 5a and b

We tested hypothesis 5a and b using the studies that reported raw scores from identical test items. This sample is smaller (111 effect sizes nested in 16 studies), and the median and weighted average duration between post and follow-up noticeably shorter (8.0 and 7.7 months, respectively).

H5a (Raw scores, identical test items) stated that for tests with identical items, the raw scores of the treatment group are stable. The first column in Table 6 shows that the intercept-only meta-regression using ES^{I} as the outcome variable resulted in a positive, large, and significant weighted average effect size. Thus, we rejected H5a. In Supplemental Material B, we show results from a sequential version of this test. The intercept was always positive, even with very short follow-up durations (shortest weighted average duration was one month). That is, there was no duration during which forgetting dominates learning in our data.

H5b (Effect sizes, identical test items) stated that if identical test items are used at post and follow-up to calculate effect sizes, then there is no fade-out. The intercept in the second column Table 6 is negative, significant, and indicates slightly more fade-out compared to the sample used to test H1-H4. Thus, we rejected H5b.

Hypotheses 6a and b

We examined H6a and b on the sample of studies that contained a comparison to an outside sample. This sample had 325 effect sizes nested in 18 studies, and the median and weighted average duration between post and follow-up test were clearly shorter than in the full sample (7 and 10.7 months, respectively). There was also less fade-out in this sample than in the samples used to test H1-H4 and H5, the weighted average was -0.12 (95% CI = [-0.19, -0.04]).

H6a (Difference outside sample, treatment) stated that the treatment group's difference to the outside sample should be stable between the post and follow-up tests. In the third column of Table 6, the coefficient on the follow-up indicator is negative and relatively large, which indicates that, relative to the outside sample, the treatment groups' scores declined between post and follow-up. However, the coefficient is not significant. We therefore failed to reject H6a.

H6b (Difference outside sample, control) stated that the difference between the control group and the outside sample should be stable between the post and follow-up tests. The coefficient on the follow-up indicator in the fourth column of Table 6 is negative, small, and not significant, that is, we failed to reject H6b.

Using only the subset of outside samples based on norm-referenced tests did not change these results much. The estimate for follow-up tests was -0.12 (95% CI = [-0.28, 0.05]) for the treatment group and -0.02 (95% CI = [-0.17, 0.14]) for the control group (based on 280 effect sizes nested in 14 studies). Excluding effect sizes based on norm-referenced tests where post and follow-up testing were conducted at different times of the year resulted in sample of 13 studies and 261 effect sizes, and a slightly smaller estimate for the treatment group, -0.09 (95% CI = [-0.25, 0.07]), and a very similar estimate for the control group, -0.05 (95% CI = [-0.22, 0.13]).

Table 6

Tests of hypotheses 5a, 5b, 6a, and 6b

	(1) H5a		(2) H5b		(3) H6a		(4) H6b	
	Coef	Df	Coef	Df	Coef	df	Coef	df
Intercept	0.40 [0.18, 0.61]	14.0	-0.18 [-0.33, -0.04]	10.7	-0.38 [-0.63, -0.13]	15.7	-0.75 [-1.0, -0.49]	16.6
Follow-up					-0.12 [-0.25, 0.01]	6.9	-0.05 [-0.18, 0.09]	6.9
Effect sizes	111		111		325		325	
Studies	16		16		18		18	
Between-study SD	0.31		0.05		0.38		0.49	
Within-study SD	0.39		0.33		0.59		0.51	
Total SD	0.50		0.33		0.70		0.70	

Note. The specifications in columns 1 (H5a) and 2 (H5b) used the sample of effect sizes based on identical test items. The specifications in columns 3 (H6a) and 4 (H6b) used the sample of effect sizes in which the treatment group (column 3) and control group (column 4) were compared to a sample of students outside of the intervention as a reference group. For hypothesis 5a, the outcome variable is the ES^{I} measure. For hypothesis 5b, the outcome variable is F, calculated using only raw scores from identical test items. For H6a and H6b, the outcome variable is ES^{norm}, calculated for the treatment and control group respectively. The df column contains the small sample adjusted Satterthwaite degrees of freedom. The terms in brackets below the coefficient estimates denotes the 95% confidence interval.

Examining Assumptions of RTM

RTM assumes that larger effect sizes are more likely to be observed at follow-up (A1) and that effect sizes are affected by random factors (A2). This section describes tests of these assumptions (see Supplemental Material E for the full description).

We examined A1 by estimating the association between having a follow-up and the weighted average effect size and the largest effect size in a subset of the sample of Dietrichson et

al. (2021). We used meta-regressions with an intercept and an indicator for having a follow-up as a single moderator, and all effect sizes or the largest effect size as the outcome variable. We found positive associations between the follow-up indicator and the weighted average effect size and the largest effect size, but only the latter association was significant.

Another potential implication of A1 and A2 is that conditional on the post-test effect size, larger effect sizes with larger standard errors are more likely to be over-estimated and should fade out more. We tested this implication using a similar CHE model as in the tests of H1-H4 with the absolute fade-out measure as the outcome variable, an intercept, and the post-test effect size, its standard error, and the interaction between the post-test effect size and the standard error as moderators. Contrary to the hypothesized implication of the assumptions, the interaction was positively associated with fade-out, albeit not significant (the standard error was positively and significantly associated with fade-out).

Conditioning on the post-test effect size means that the standard error is fully determined by the sample size. Thus, a potential explanation may be that small-study effect sizes fade-out less. We discuss reasons why this might be the case, and provide evidence against one such explanation–that smaller studies may, by chance, be more likely to obtain unbalanced treatment and control groups before the intervention–in Supplemental Material E.

Examining the Main Assumption of the Increasing Variance Mechanism

The increasing variance mechanism assumes that increasing student heterogeneity over time increases the variance of test scores. We tested whether the standard deviation is larger at follow-up than at post-test in the sample of identical tests. We used the sample of identical tests to avoid confounding increasing variance due to increased student heterogeneity with increasing variance due to follow-up tests being broader in content or containing more test items. We found

no evidence that the test variances increase over time in this sample (for details, see Supplemental Material F).

Sensitivity Analyses

In this section, we describe sensitivity analyses that concern more than one hypothesis test. We tested sensitivity to sample restrictions based on post-test effect sizes, using different values of ρ, using only identical samples, using a CE-RVE model, including additional moderators (the grade and the duration of the intervention, and whether the implementers were school personnel), using only within-study variation in the estimations, using a relative fade-out measure, and restricting the sample to RCTs of tutoring interventions. We report the full description of the tests and the results in the Supplemental Material D, and focus here on the cases when a hypothesis test was not entirely robust.

H1 (Sample composition) was reasonably robust to all sensitivity tests. A partial exception was that the weighted average effect size in identical samples was negative but not significant when we used the full sample. Restricting the sample to treatments with larger post-test effect sizes yielded instead a more negative and more significant effect size. As this pattern of results was in line with there being a lower bound for fade-out at zero, we believe that that the rejection of H1 was robust.

H2 (Follow-up duration) was not robust to restricting the sample to treatments with larger post-test effect sizes. The coefficient on the indicator for post to follow-up durations longer than 17 months became progressively more negative, the more we restricted the sample and was significant for ES > 0.2 and > 0.3. Furthermore, both coefficients were negative and significant when we used the relative fade-out measure.

For H5a (Raw scores, identical test items), the weighted average effect size remained positive but turned insignificant when we restricted the sample to identical samples. However, the mechanisms either predicted a negative and significant effect size (skill decay) or were compatible with both a stable and a positive effect size (compensatory catch-up and the treatment-induced catch-up mechanisms). Thus, the insignificant estimate did not change the support for or against the mechanisms.

H5b (Effect sizes, identical test items) turned insignificant when we restricted the sample to identical tests and when we restricted the sample to RCTs of tutoring interventions. However, in both cases, the weighted average fade-out was actually more negative than in the primary analysis, and the insignificance may have been related to the low statistical power in this subset of the data (7 and 11 studies, respectively). The coefficient was also negative, indicating fade-out, but not significant when we used the relative fade-out measure.

We found a negative and significant estimate when using the CE-RVE model to test H6a (Difference outside sample, treatment). We also found a negative and significant estimate of the follow-up indicator when testing H6b (Difference outside sample, control) on identical samples. However, the small number of identical samples (6 studies) should temper the conclusions drawn from this sensitivity analysis.

Discussion

Table 7 summarizes the outcomes of the hypothesis tests by mechanism. For each hypothesis, the table displays the sign of the $\hat{\beta}$ (0 indicates that we failed to reject the hypothesis in the primary analysis) and our assessments of the strength of the evidence (dark green or red indicate robust evidence, and light green and light red not robust evidence). We discuss our assessments, first by hypothesis and then by mechanism. In the latter discussion, we also discuss the

tests of assumptions and compare our results to evidence from primary studies. In the penultimate section, we discuss more general limitations of the review. Lastly, we offer concluding remarks and directions for future studies.

Table 7

Summary of hypothesis tests by mechanism

Mechanism	H1	H2	H3	H4	H5a	H5b	H6a	H6b
Skill-based mechanisms								
Skill decay		0	-		+	-	0	0
Compensatory catch-up		0	-	0	+	-	0	0
Treatment induced (t)		0	-		+	-	0	0
Treatment induced (c)		0	-		+	-	0	0
Statistical and measurement mechanisms								
Sample composition	-							0
Regression to the mean		0	-			-	0	0
Increasing variance		0				-	0	0
Vertical scaling of tests						-	0	0

Note. For each hypothesis, the table displays the sign of the $\hat{\beta}$: + for positive, - for negative, and 0 if we failed to reject the hypothesis. Blank cells indicate that the mechanism was compatible with any outcome of the hypothesis test. The colors of the cells indicate whether the outcome of the hypothesis test was in line with the prediction of the mechanism. Dark green indicates that the test was robustly in line with the mechanism's prediction, light green that it was in line but not robust, dark red that the test was not in line and robust, light red that it was not in line but not robust.

Interpreting the Hypothesis Tests

There are threats to interpreting the hypothesis tests as evidence for or against the fade-out mechanisms. There may be omitted variables that influenced both the variables included in

our models and the outcome measure; that is, there is risk of unmeasured confounding. The variables we included in the meta-regressions may not have perfectly captured the hypothesized relation with fade-out (i.e., measurement error), and not all tests were robust. Below, we discuss the interpretation of each hypothesis test in the light of these problems.

We rejected H1 (Sample composition), the hypothesis that there is no or less fade-out in samples in which the sample composition did not change between the post and follow-up test. Because we found fade-out also in these identical samples, the sample composition mechanism was not a major explanation of fade-out in our data. The sensitivity analyses did not give us reason to change this conclusion. However, in some specifications we found slightly less fade-out in identical samples than in non-identical samples, although the difference was not significant. That is, changes to the sample composition may have influenced fade-out to some degree.

In the primary analysis, we failed to reject H2 (Follow-up duration), which stated that the fade-out of effect sizes is unrelated to the duration between post and follow-up test. We found relatively large, negative but insignificant estimates of the association between duration and fade-out. Relatively small amounts of unmeasured confounding or measurement error–e.g., our two indicator-specification may not have captured the association between duration and fade-out well enough–could have changed the results. Furthermore, we found more negative and sometimes significant estimates in the sensitivity analyses as well as evidence of selection also of follow-up effect sizes, which may have biased the primary analysis coefficients toward null. While not robust, we interpreted the results of this hypothesis test as being more in line with mechanisms predicting more fade-out over time.

We rejected H3 (Post-test effect size), which stated that the fade-out of effect sizes was unrelated to the post-test effect size. The association was strongly negative and robust across all

sensitivity analyses. Because the association was strong also when we restricted the sample to treatments with relatively large post-test effect sizes, the result seems unlikely to be a mechanical effect of a lower bound of fade-out at zero. Although the estimate may be influenced by study and effect size characteristics that were not included in our models, it seems unlikely that it would change the direction of the association. We interpreted the collected results as robust evidence in favor of mechanisms predicting more fade-out of larger post-test effect sizes.

We could not reject H4 (Overlapping test domains), which stated that the degree of overlap between the domains tested at post and follow-up is unrelated to fade-out. This result was robust across all sensitivity tests. The confidence intervals were always wide, the sign of the coefficient on identical tests was always positive, indicating less fade-out, and the magnitude was mostly small. It was however noticeably larger when we only used within-study variation in the estimation, which may indicate that there was unmeasured study-level confounding in the other estimations. Furthermore, the overlap of test content was relatively high also among non-identical tests. We may therefore not have had enough statistical power to capture the relatively small differences in the probability of skill transfer between the types of tests. We interpreted the results as being more in line with there being less fade-out when test content was more overlapping, but the evidence was clearly not robust.

We rejected H5a (Raw scores, identical test items) as we found that the treatment group significantly increased their raw scores between post and follow-up when tested on identical test items. The treatment group increased their raw scores also in an analysis confined to identical test items and identical samples, albeit less so. A potential source of unmeasured confounding would be if the treatment group received further interventions between the post and follow-up test and the control group did not. Only one of the studies in our sample mentioned the potential

for such interventions (Svensson et al., 2021), and we rejected H5a also when this study was excluded from the sample. Except for one study, we received confirmation from the authors that tests were identical. Our results were not sensitive to excluding the exception or studies with additional concerns (e.g., start and stop-rules). That is, the scope for measurement error ought to be small. Thus, we were unable to find another explanation of the rejection of H5a than that the treatment group's skills on average did not decay.

We rejected H5b (Effect sizes, identical test items) that there was no fade-out of effect sizes calculated using the raw scores from identical test items. This result was not robust to restricting the sample to identical samples and to RCTs of tutoring interventions, although the weighted average fade-out was actually more negative in these two samples than in the primary analysis. It was also not significant when we used the relative fade-out measure. As discussed, we believe the relative fade-out measure has more problems than the absolute measure. The advantage of the relative measure was primarily that the zero lower bound was less of a concern. Because the H5b test was not sensitive to including only treatments with large average effects, the zero lower bound ought to be less of a problem for this hypothesis test. We therefore interpreted the results as robust evidence that there was substantial fade-out also with identical test items.

In the primary analysis, we failed to reject H6a (Difference outside sample, treatment) and H6b (Difference outside sample, control), which concerned comparisons of the post to follow-up development of an outside sample to the treatment and control group, respectively. However, the treatment group's decline in relation to the outside sample was not small and it was significant when we used the CE-RVE model. We found a negative and significant development in the control group compared to the outside sample when only using identical samples, although

the small sample size made this result uncertain. The confidence interval was relatively wide in the primary analysis of H6b as well.

These two hypotheses rested on the strong assumption that the development, on average, would have been parallel to that of the outside sample, had there not been a study. One reason to believe that the assumption did not hold strictly is that test variances may increase over time, which would imply that the relative differences between outside samples and treatment and control groups would become smaller over time as measured by our effect sizes. However, we found no evidence of increasing variances and earlier studies indicated that any increase ought to be very small over time periods such as the duration between post and follow-up tests in our study (Andersen et al., 2022; Cascio & Staiger, 2012; Wan et al., 2021).

Another reason may be that skill differences between at-risk students, who make up most of our sample, and not-at-risk students, who are included in the outside samples, increase over time. However, previous studies have found both increases, decreases, and no changes over the school years (e.g., Caro et al., 2009; Geary et al., 2012; Halle et al., 2012; Heckman, 2006; Jordan et al., 2003; Lipsey et al., 2012; Pace et al., 2019; Sirin, 2005; von Hippel et al., 2018). As we discuss in detail in Supplemental Material G, most findings indicate that the changes would be small over the time elapsed from post to follow-up in our sample.

Taken together, we believe that the tests of H6a were not robust but more in line with mechanisms predicting a decline in the treatment groups' score relative to the outside sample than those that predicted a stable score difference. Similarly, the tests of H6b were not robust but more in line with mechanisms predicting a stable score difference than an increase in the control groups' relative score.

Implications for Skill Decay

Predictions of skill decay were in line with the robust rejections of H3 (Post-test effect size) and H5b (Effect sizes, identical test items). The results regarding H2 (Follow-up duration), H6b (Difference outside sample, control), and H6a (Difference outside sample, treatment) were also more in line with the predictions of skill decay than not but were not robust. However, the test of H5a (Raw scores, identical test items), which directly tested the core prediction of skill decay, showed the opposite of decay. Because we believed the scope of unmeasured confounding and measurement error to be small for this hypothesis and because other mechanisms also were in line with the tests supporting skill decay, we interpreted the rejection of H5a as relatively strong evidence that skill decay was not a prominent explanation of fade-out in our data.

These results are in line with primary studies that have examined whether the skills of the treatment group decay after post-test (; Bailey, Fuchs, et al., 2020; Bøg et al., 2021; Carr et al., 2024; Kang et al., 2019). Kang et al. (2019) presented a detailed test item-level analysis in the context of an early childhood mathematics intervention and found some evidence of the treatment group forgetting their newly acquired skills. However, forgetting only explained a small proportion of fade-out. The other studies found that the treatment group on average increased their scores on all tests with comparable scores across time.[14]

These results do not rule out that forgetting explains a small proportion of fade-out, as found by Kang et al. (2019), or that some interventions may have only short-term effects on skills. The concerns about skill decay were not unfounded in our sample: eight studies included negative values on at least one effect size used in the analysis of H5a. Furthermore, that we

[14] In a similar vein, Protzko (2016) found no support for fade-out of IQ scores being caused by intervention effects only affecting test-taking skills and then being forgotten. Instead, the results supported an effect on intelligence.

found little support for skill decay on average does not imply that interventions should not make use of best practices to help students remember what they learn (see e.g., Vlach (2019) for a discussion). Indeed, one reason why skills on average did not seem to decay in our data may be that interventions included such best practices. But the results should make us question skill decay–an absolute post-intervention decline in skills–as a major explanation of fade-out.

Implications for Compensatory Catch-Up

The tests of H3 (Post-test effect size), H5a (Raw scores, identical test items), and H5a (Effect sizes, identical test items) provided robust support for the compensatory catch-up mechanism. Although not providing robust support, H2 (Follow-up duration), H4 (Overlapping test domains), H6a (Difference outside sample, treatment), and H6b (Difference outside sample, control) were more in line with the predictions of the compensatory catch-up than not. Thus, no hypothesis test provided robust evidence against compensatory catch-up.

The empirical evidence for compensatory catch-up is mixed in the previous literature. For example, Bailey et al. (2017) and Burchinal et al. (2022) found that skills that are malleable, fundamental, and would not have developed without intervention were less likely to fade out. Ansari et al. (2020) and McCormick et al. (2021) found that unconstrained skills were associated with less fade-out than constrained skills. On the other hand, the meta-analysis of Rosengarten et al. (2024) found, if anything, more fade-out of unconstrained skills and Li et al. (2020) found no difference in fade-out between achievement tests and cognitive skills tests in their meta-analysis, where the latter probably tests more unconstrained skills.

Bailey, Jenkins, et al. (2020) reviewed studies of preschool interventions that also measured later school quality but found little evidence for the sustaining environments hypothesis. Burchinal et al. (2022) similarly found no support for the sustaining environments hypothesis in

a recent study. However, neither the studies included in the meta-analysis of Bailey, Jenkins, et al. (2020) nor the analysis of Burchinal et al. (2022) made use of random or quasi-experimental variation in classroom or school quality.[15] The relation between classroom/school quality and the persistence of effects may therefore be confounded in these studies or biased due to measurement error. Our results support treating this mechanism as one of the major candidate explanations of fade-out.

Implications for Treatment-Induced (t)

The results regarding H3 (Post-test effect size), H5a (Raw scores, identical test items), and H5b (Effect sizes, identical test items) provided robust support for treatment-induced (t). Although not robust, H2 (Follow-up duration), H6a (Difference outside sample, treatment), and H6b (Difference outside sample, control) were more in line with the predictions of treatment-induced (t) than not. Thus, no hypothesis test provided robust evidence against this mechanism.

To the best of our knowledge, no primary study of school interventions has fully examined this mechanism. There is evidence that parents react to other types of interventions in ways that would thwart the improvements of the type of interventions examined in this study (Currie & Thomas, 2000; Das et al., 2013; Fredriksson et al., 2024; Gensowski et al., 2020; Greaves et al., 2023; Pop-Eleches & Urquiola, 2013).[16] Similarly, schools may reduce their help to the treatment group after a successful intervention (Blachman et al., 2014; Fredriksson et al., 2024). But we are not aware of studies that have examined both these entities' reactions to a successful in-

[15] Jenkins et al. (2018) found that targeted teacher professional support between kindergarten and first grade after a preschool intervention mitigated fadeout. However, the mitigating effect was not mediated by measures of classroom quality.

[16] As mentioned in the Treatment-induced catch-up-section, this pattern is not universal and there are examples of preschool interventions that change parental behavior in a way that ought to increase the chances of growing effects instead of fade-out (Gelber & Isen, 2013; Kitchens et al., 2020; Reynolds & Ou, 2011; Watts et al., 2020).

tervention, and the potential ways they may react go beyond the examples examined in the literature. Our results provided reasons to include surveys of parents and schools in longitudinal studies of interventions and to regard treatment-induced (t) as one of the major candidate explanations of fade-out.

Implications for Treatment-Induced (c)

The results regarding H3 (Post-test effect size), H5a (Raw scores, identical test items), and H5b (Effect sizes, identical test items) provided robust support for this mechanism. The test of H2 (Follow-up duration) was more in line with the predictions of treatment-induced (c) than not, whereas H6a (Difference outside sample, treatment) and H6b (Difference outside sample, control) were not. If the control group received extra help because the treatment group received a successful intervention, as posited by treatment-induced (c), then we expected an improved score for the control group relative to outside samples and a stable score for the treatment group. We could not reject stable differences in the primary analysis, but did reject them in some sensitivity analyses and all estimates indicated declining relative scores for both groups. For H6b, the sign was thus opposite to the prediction of treatment-induced (c), although the confidence interval often included sizeable positive estimates.

To the best of our knowledge, there is no evidence from primary studies that parents and schools increase their efforts to help the control group because they participated in a study in which an intervention improved the treatment group's skills. This does not mean that the control group does not get any help. Indeed, one explanation for the patterns we see is that the treatment-as-usual involves special educational services or, more generally, help that the control group receives and the treatment group does not. It is just that the evidence that the control group students get *extra* help, because they were the control group, is not strong. Although examining

treatment-induced (c) further would be relevant on the basis of our results, we regard it as a less likely explanation of fade-out than treatment-induced (t).

Implications for Sample Composition

H1 (Sample composition) tested the sample composition mechanism directly. We found fade-out also in identical samples and the hypothesis was rejected. Furthermore, H6b (Difference outside sample, control) provided some, albeit weaker, evidence against this mechanism. However, some specifications showed less fade-out in identical samples than in non-identical samples. Because these differences were not consistent across specifications and never significant, changes to the sample composition were unlikely to be a major explanation of fade-out in our data.

One plausible reason why changes to the sample composition did not seem to cause fade-out was that the difference between treatment and control group attrition rates was low (on average 1.2%, see Table 2). Another reason was that in around 50% of the studies with non-identical samples, we were able to use baseline-adjusted effect estimates as the basis for effect sizes and fade-out measures. Such adjustments ought to mitigate sample composition problems. Both these features of our data strengthened the interpretation that sample composition was an unlikely cause of fade-out.

We are not aware of other studies that examine whether changes in the sample compositions on average favor the treatment or the control group. That is, although our results made sample composition changes seem like an unlikely explanation of fade-out in our data, information about such changes needs to be included in any examination of fade-out.

Implications for Regression to the Mean

The results regarding H3 (Post-test effect size) and H5b (Effect sizes, identical test items) provided support for RTM and were robust. That we failed to reject H2 (Follow-up duration) was in line with RTM, but this result was not robust across specifications and some versions of the test did not align with the implications of RTM. We failed to reject both H6a (Difference outside sample, treatment) and H6b (Difference outside sample, control), although we found significant and negative score differences for the former in sensitivity analyses, in line with RTM.

We tested the two assumptions that give rise to RTM–that larger post-test effect sizes increase the probability of a follow-up and that post-test effect sizes are affected by random factors–and found mixed support. Overall, some degree of effect size-based selection seems plausible. Rather than asking whether RTM explains fade-out, we might ask how much it could explain. To this end, we calculated how much fade-out could potentially be explained by RTM (see Supplemental Material E). The calculations rest on several assumptions and approximations so the exact numbers should be interpreted with caution. Nevertheless, even with an implausibly high cutoff (the empirical mean) for when effect sizes were followed-up, RTM explained far from all fade-out. With a cutoff of zero, RTM explained a small but non-trivial proportion of fade-out (around 20%).

Implications for Increasing Variance

The results regarding H5b (Effect sizes, identical test items) were in line with the increasing variance mechanism. However, we found no evidence that the standard deviations were larger at follow-up than at post-test–a necessary condition for the increasing variance mechanism–in a sample including all identical tests, as well as in the same sample used to test H5b. Be-

cause other mechanisms were also in line with the results regarding H5b, the fade-out was therefore unlikely to be caused by the increasing variance mechanism. The test of H2 was in the predicted direction, whereas H6a and H6b both showed the wrong sign.

The lack of support for the increasing variance mechanism was in line with previous primary studies. Both Cascio and Staiger (2012) and Wan et al. (2021) found that increasing variance explained some fade-out. In contrast, Andersen et al. (2022) found that the variance decreased over time. However, the estimated magnitudes of increasing/decreasing variance were small in all three studies, and would imply a negligible effect on fade-out over the time elapsed between post to follow-up in almost all of our included studies. Thus, increasing variance is a factor that should be considered when assessing fade-out in the long run. However, with post to follow-up duration as in our sample, this mechanism seemed negligible.

Implications for Vertical Scaling of Tests

H5b (Effect sizes, identical test items) directly concerned the vertical scaling of tests mechanism. We found significant fade-out of effect sizes when they were calculated based on raw scores from identical test items–for which vertical scaling is not an issue–which provided robust evidence against this mechanism. The mechanism predicted a rejection of either H6a (Difference outside sample, treatment) or H6b (Difference outside sample, control), or both. We failed to reject both H6a and H6b in most of our analyses, providing mixed support for the vertical scaling of tests mechanism.

Wan et al. (2021) showed that vertical scaling of tests can change the degree of fade-out quite dramatically, which in some of their analyses lead to an overestimation of fade-out. Vertical scaling issues may cause either growth or fade-out depending on the context. Although our results did not indicate that the mechanism caused fade-out in our data, more research is needed.

Summary and Agreements with Earlier Reviews

We interpreted our results as being most compatible with more than one mechanism causing effect sizes of school interventions to fade out. However, not all surveyed mechanisms were equally supported: we found robust evidence against the skill decay, increasing variance, sample composition, and vertical scaling of tests mechanisms. Further evidence against the latter three mechanisms was that they did not feature the post-test effect size in their explanations of fade-out, which, because the post-test effect size explained all between-study heterogeneity, restricted their explanatory power.

We found no robust evidence against compensatory catch-up, treatment-induced (t), treatment-induced (c), and RTM. However, we found more support for the first two than the latter two mechanisms and we found mixed support for the RTM assumptions. In Protzko (2015), fade-out was caused by a relative decline of the IQ scores of the treatment group. Compensatory catch-up, treatment-induced (t), and RTM are compatible with the results in Protzko (2015; as are the skill decay and vertical scaling of tests mechanisms). Against the implications of treatment-induced (c), sample composition, and increased test variance mechanisms, Protzko (2015) found fade-out because of a relative decrease of the treatment group's scores while the control group's scores were stable.

Bailey, Duncan, et al. (2020) argued that fade-out depends on the types of skills targeted by the intervention, and on complementarities between interventions and subsequent environmental affordances. Abenavoli (2019) found evidence of family and school support mediating the longer-term effects. These findings were in line with compensatory catch-up and treatment-induced catch-up. Bailey, Duncan, et al. (2020) also concluded that fade-out may to some extent be

a measurement artifact, but it is unlikely to fully be explained by measurement issues. Our results agreed. In fact, we found robust evidence against the two pertinent measurement issues in our sample. Thus, measurement issues may be a negligible source of fade-out, given the type of tests used and the short time elapsed between post-test and follow-up, as in our data.

In sum, there were no results in earlier reviews that contradicted ours.[17] On the contrary, the mechanisms that were most supported in our analyses–compensatory catch-up and treatment-induced (t)–were also supported by Protzko (2015), Abenavoli (2019), and Bailey, Duncan, et al. (2020). It is also noteworthy that the results in Protzko (2015) provided evidence against the treatment-induced (c) mechanism. None of these reviews examined RTM quantitatively.

Limitations

Our study has some important limitations. The analysis was not pre-registered. As at least one of the authors was also a co-author of the systematic reviews from which we drew most of our studies, we were familiar with the data before starting this project. Thus, pre-registering the analyses before seeing the data was not an option. Our aim was exploratory, but we hope that the survey and discussion of mechanisms and the hypothesis tests may serve as an inspiration for future confirmatory primary studies and meta-analyses.

The current sample was not collected with the intention of finding all studies of educational interventions with a post and follow-up measurement (i.e., studies in which the measurement of fade-out is possible). However, as detailed in the section The Present Study, the sample was the result of a large search effort.

[17] As shown in Supplemental Material D, in contrast to Li et al. (2020), we found no association with the starting time of the intervention (the grade in our case, the age of preschool children in Li et al. 2020) and the duration of the intervention. However, as we discuss, these variables were ambiguously related to the fade-out mechanisms in our context. Furthermore, Li et al. (2020) did not include the post-test effect size in their meta-analysis, which explained all between-study heterogeneity in our analysis.

The sample constrained how much the results can be generalized to other school interventions, and, more so, to interventions in preschool or tertiary education. In particular, one feature of our sample that may affect both the degree of fade-out and the importance of certain mechanisms was the high prevalence of within-school assignment of treatment. School-level assignment of treatment seems less likely to change the schools' internal prioritization of which students should receive extra help and may make students' relative achievement less visible to parents than within-school assignment. Therefore, school-level assignment may weaken the side effects of a successful intervention; this was how treatment-induced (t) and (c) explained fade-out. School-level assignment may furthermore imply that more students at the same schools are treated, which may create stronger peer effects and a more sustaining environment.[18] Thus, it may weaken the compensatory catch-up mechanism.

Our review examined fade-out of effect sizes based on standardized tests of mathematics and reading, in most cases with a relatively short follow-up period. There are examples of studies finding very long-term effects of school interventions on measures such as educational attainment, earnings, and health (e.g., Chetty et al., 2011; Fischer et al., 2021; Fredriksson et al., 2013; Meghir & Palme, 2005). These measures have in common that they depend on a much broader set of skills than tests of reading and mathematics (Dietrichson, Kristiansen, & Viinholt, 2020).

Some evidence from preschool interventions indicate that effects may be more lasting on skill measures that involve social-emotional (or non-cognitive) skills (Abenavoli, 2019; Bailey et al., 2017; Deming, 2009; Heckman et al., 2013; McCoy et al., 2017; Wolf et al., 2019), which we did not examine because our data only included tests of reading and math. However, the meta-

[18] Bai et al. (2020) hypothesized that two preschool programs, which showed increasing effects over time, might have done so because they affected many children and thus may have changed the social ecology and peer culture into a sustaining environment.

analysis by Hart et al. (2024) found that effects on social-emotional skills fade out at least as fast as effects on cognitive skills. Pages et al. (2023) found a larger mediation effect for cognitive skills than for social, emotional, and behavioral skills on completed years of schooling for the Abecedarian and Head Start programs.

Hypotheses 6a and b compared the development of the treatment and control groups with samples outside of the intervention. The hypotheses were derived under the strong assumption that the development of the treatment and control group would have been parallel to the outside sample, had they not participated in the intervention. Because the outside samples contained normally achieving students, which most of our intervention samples did not, there is uncertainty of whether this assumption holds. Using outside samples of students that are as similar as possible to the treatment and control group would have been preferable, but no study in our sample included such data. In the Supplemental Material G, we probed the sensitivity of our results to violations of this assumption by using estimates of the trends from the earlier literature.

Conclusions and Directions for Future Research

We surveyed and examined empirically mechanisms of fade-out in school interventions. In line with two previous reviews (Abenavoli, 2019; Bailey, Duncan, et al., 2020), our results indicated that fade-out was not caused by a single mechanism. We found most support for mechanisms that explain fade-out by the control group learning the skills taught in the intervention during regular post-intervention instruction (compensatory catch-up) and the treatment having negative side effects on the post-intervention investments received by the treatment group (treatment-induced (t)). We also found some support for and no robust evidence against mechanisms explaining fade-out by studies being selected for follow-up because post-test effect sizes were large (RTM), and the control group students receiving extra help because they were in the control

group (treatment-induced (c)). The first three mechanisms were also compatible with the development of IQ scores after preschool interventions in the meta-analysis of Protzko (2015), which the treatment-induced (c) mechanism was not.

We found evidence against fade-out being caused by the treatment group forgetting more of their skills after the intervention than they learn (skill decay), changes to the sample composition, increasing test variance, and differences in the vertical scaling of tests. Of these, skill decay and vertical scaling of tests were compatible with the pattern of results in Protzko (2015). Thus, where we could compare our results, they were largely in agreement with earlier meta-analyses. It is furthermore noteworthy that RTM was unlikely to explain a large share of fade-out in our data. Thus, most fade-out is likely not explained by an inability to properly measure and estimate effects but by the treated students' skills growing more slowly than the control students' after the intervention.

Based on our results and those of the earlier literature, compensatory catch-up and treatment-induced (t) seem like the most likely explanations of fade-out. Our results also indicated that RTM, which has not been tested in other studies, explains some share of fade-out. These three mechanisms have quite different implications for both the returns to interventions and theories of child development. For the returns to interventions, treatment-induced (t) represents a possible best-case scenario and RTM a possible worst-case scenario. If treatment reduces parental and school investments, the saved costs of these investments would improve the returns to intervention–potentially to the point where an intervention becomes a good investment even if the effects on skills were to fade out completely. In contrast, RTM implies that the short-term effects of interventions with follow-ups were over-estimated.

Compensatory catch-up would also imply relatively low returns. How this reflects on the intervention may depend on the cause of the catch-up. If the cause is a failure of the intervention to target the right type of skills, then the effects were "doomed" to fade out. If the fade-out is instead caused by a failure of schools and parents to help the treatment group build on the skills they learnt in the intervention, then the intervention created a potential for high returns in the long run, as well as for further learning (Bransford & Schwartz, 1999; Clements et al., 2013). That is, it is not changes to the intervention that are needed to boost returns, but changes to the subsequent school or family environment.

Regardless of the reasons, compensatory catch-up is incompatible with theories predicting unconditional Matthew effects of successful interventions. In contrast, both treatment-induced (t) and RTM are potentially compatible with such theories. As mentioned, treatment-induced (c), changes to the sample composition, and all mechanisms based on measurement issues are also compatible with an underlying skill development process that exhibits Matthew effects. The requirement that the Matthew effects be smaller than the combined effect of these mechanisms suggests that non-trivial Matthew effects could coexist with fade-out.

Further research about fade-out mechanisms is clearly needed. We believe both new meta-analyses and new primary studies are necessary to learn more. Including a larger set of education interventions would improve the statistical power when testing our, or other, hypotheses. Including a broader set of outcomes in the meta-analysis would enable the examination of the hypothesis that fade-out depends on the type of skills and outcome measures. Although such hypotheses could also be tested in primary studies, meta-analysis is needed to examine the generality of the phenomenon.

More primary studies that examine the degree to which compensatory catch-up and treatment-induced catch-up are at work simultaneously and heterogeneously across students would be very valuable. Because compensatory catch-up and treatment-induced (t) predict that fade-out happens due to the treatment group's skills growing slower than the control group, it was difficult to develop hypotheses about the effect size distribution that directly tested them against one another. Information on how schools and parents react to successful interventions would be one way to separate between them. As suggested by Gelber and Isen (2013), analyses of fade-out should include measures of parents' and schools' post-intervention behavior.

However, we also want to acknowledge that the amount of information required to evaluate several mechanisms simultaneously in primary studies is quite formidable. For example, no study in our sample included enough information to evaluate whether the treatment-induced catch-up mechanisms were important influences on fade-out. Furthermore, a complete evaluation of fade-out is not possible in any primary study due to the impossibility of knowing the extent to which the effects are affected by random factors outside of the intervention (i.e., to examine RTM). Thus, more studies examining one mechanism at a time would still be valuable, as would studies with more than one follow-up occasion.

The inability of any primary intervention study to examine all potential fade-out mechanisms and the incomplete knowledge of these mechanisms in the literature should moderate the policy conclusions drawn from long-term evaluations of single interventions. While long-term evaluations of interventions are necessary to gain more knowledge about fade-out (Bailey et al., 2024; Andersen et al., 2022), without enough information about mechanisms, such evaluations

are not necessarily informative about whether an intervention should be scaled up or discontinued.[19]

[19] As an example, the effects of being randomized to the Tennesee Voluntary Prekindergarten Program were positive in the short-term but faded out or turned negative in third and sixth grade (Durkin et al., 2022; Lipsey et al., 2018; Watts, Duncan et al., 2019). These results have been much debated (see e.g., Whitehurst, 2018 for an account of this debate). However, the studies did not include information about how, e.g., parents reacted to the interventions. We believe it is unclear what caused the effects to fade-out or become negative and therefore unclear what policy conclusions that should be drawn.

References

References marked with an asterisk indicate records included in the meta-analysis.

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies*, *72*(1), 1-19. https://doi.org/10.1111/0034-6527.00321

Abenavoli, R. M. (2019). The mechanisms and moderators of "fade-out": Towards understanding why the skills of early childhood program participants converge over time with the skills of other children. *Psychological Bulletin*, *145*(12), 1103–1127. https://doi.org/10.1037/bul0000212

*Alan, S., Boneva, T., & Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *Quarterly Journal of Economics*, *134*(3), 1121–1162. https://doi.org/10.1093/qje/qjz006

Andersen, S. C., Bodilsen, S. T., Houmark, M. A., & Nielsen, H. S. (2022). *Fade-out of educational interventions: Statistical and substantive sources*. (CESifo Working Paper No. 10094). CESifo, Munich.

Anderson, M. C., & Hulbert, J. C. (2021). Active forgetting: Adaptation of memory by prefrontal control. *Annual Review of Psychology*, *72*, 1–36. https://doi.org/10.1146/annurev-psych-072720-094140

Ansari, A., Pianta, R. C., Whittaker, J. V., Vitiello, V. E., & Ruzek, E. A. (2020). Persistence and convergence: The end of kindergarten outcomes of pre-k graduates and their nonattending peers. *Developmental Psychology*, *56*(11), 2027–2039. https://doi.org/10.1037/dev0001115

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Bai, Y., Ladd, H. F., Muschkin, C. G., & Dodge, K. A. (2020). Long-term effects of early childhood programs through eighth grade: Do the effects fade out or grow? *Children and Youth Services Review*, *112*, 104890. https://doi.org/10.1016/j.childyouth.2020.104890

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest*, *21*(2), 55–97. https://doi.org/10.1177/1529100620915848

Bailey, D. H., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*(1), 7–39. https://doi.org/10.1080/19345747.2016.1232459

Bailey, D. H., Duncan, G. J., Watts, T., Clements, D. H., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *American Psychologist*, *73*(1), 81–94. https://doi.org/10.1037/amp0000146

*Bailey, D. H., Fuchs, L. S., Gilbert, J. K., Geary, D. C., & Fuchs, D. (2020). Prevention: Necessary but insufficient? A 2-year follow-up of an effective first-grade mathematics intervention. *Child Development*, *91*(2), 382–400. https://doi.org/10.1111/cdev.13175

Bailey, D. H., Jenkins, J. M., & Alvarez-Vargas, D. (2020). Complementarities between early educational intervention and later educational quality? A systematic review of the sustaining

environments hypothesis. *Developmental Review*, *56*(February), 100910. https://doi.org/10.1016/j.dr.2020.100910

Bailey, D. H., Watts, T. W., Hart, E. R., & Yu, M. J. (2024). Learning about development from interventions. *Annual Review of Developmental Psychology*, *6*(1), 251-272. https://doi.org/10.1146/annurev-devpsych-010923-103044

Barnett, A. G., Van Der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637. https://doi.org/10.1037//0033-2909.128.4.612

*Blachman, B. A., Schatschneider, C., Fletcher, J. M., Francis, D. J., Clonan, S. M., Shaywitz, B. A., & Shaywitz, S. E. (2004). Effects of intensive reading remediation for second and third graders and a 1-year follow-up. *Journal of Educational Psychology*, *96*(3), 444–461. https://doi.org/10.1037/0022-0663.96.3.444

*Blachman, B. A., Schatschneider, C., Fletcher, J. M., Murray, M. S., Munger, K. A., & Vaughn, M. G. (2014). Intensive reading remediation in grade 2 or 3: Are there effects a decade later? *Journal of Educational Psychology*, *106*(1), 46–57. https://doi.org/10.1037/a0033663

Bond, T. N., & Lang, K. (2013). The evolution of the black-white test score gap in grades k–3: The fragility of results. *Review of Economics and Statistics*, *95*(5), 1468–1479. https://doi.org/10.3386/w17960

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons. https://doi.org/10.1002/9780470743386

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications (vol. 24). *Review of Research in Education*, *24*, 66–100. https://doi.org/10.2307/1167267

Burchinal, M., Foster, T., Garber, K., Cohen-Vogel, L., Bratsch-Hines, M., & Peisner-Feinberg, E. (2022). Examining three hypotheses for pre-kindergarten fade-out. *Developmental Psychology*, *58*(3), 453–469. https://doi.org/10.1037/dev0001302

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

Bøg, M., Dietrichson, J., & Isaksson, A. A. (2021). A multi-sensory tutoring program for students at risk of reading difficulties: Evidence from a randomized field experiment. *Journal of Educational Research*, *114*(3), 233–251. https://doi.org/10.1080/00220671.2021.1902254

Camilli, G., Vargas, S., Ryan, S., & Barnett, W. S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record*, *112*(3), 579–620. https://doi.org/10.1177/016146811011200303

Campbell, D. T., & Frey, P. W. (1970). The implications of learning theory for the fade-out gains from compensatory education. *Compensatory Education: A National Debate*, *3*, 455–463.

Caro, D. H., McDonald, J. T., & Willms, J. D. (2009). Socio-economic status and academic achievement trajectories from childhood to adolescence. *Canadian Journal of Education/Revue Canadienne de l'éducation*, *32*(3), 558–590.

Carr, R. C., Jenkins, J. M., Watts, T. W., Peisner-Feinberg, E. S., & Dodge, K. A. (2024). Investigating if high-quality kindergarten teachers sustain the pre-K boost to children's emergent literacy skill development in North Carolina. *Child Development*, online advance access. https://doi.org/10.1111/cdev.14076

Cascio, E. U., & Staiger, D. O. (2012). *Knowledge, tests, and fadeout in educational interventions* (No. 18038). National Bureau of Economic Research. https://doi.org/10.3386/w18038

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Chapman, J. W., Tunmer, W. E., & Prochnow, J. E. (2000). Early reading-related skills and performance, reading self-concept, and the development of academic self-concept: A longitudinal study. *Journal of Educational Psychology*, *92*(4), 703–708. https://doi.org/10.1037/0022-0663.92.4.703

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from project STAR. *Quarterly Journal of Economics*, *126*(4), 1593–1660. https://doi.org/10.1093/qje/qjr041

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292. https://doi.org/10.3102/0013189X16656615

*Cirino, P. T., Vaughn, S., Linan-Thompson, S., Cardenas-Hagan, E., Fletcher, J. M., & Francis, D. J. (2009). One-year follow-up outcomes of Spanish and English interventions for English language learners at risk for reading problems. *American Educational Research Journal*, *46*(3), 744–781. https://doi.org/10.3102/0002831208330214

*Clarke, B., Doabler, C., Smolkowski, K., Kurtz Nelson, E., 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*(4), 607–634. https://doi.org/10.1080/19345747.2015.1116034

*Clarke, B., Doabler, C. T., Sutherland, M., Kosty, D., Turtura, J., & Smolkowski, K. (2023). Examining the impact of a first grade whole number intervention by group size. *Journal of Research on Educational Effectiveness*, *16*(2), 326–349. https://doi.org/10.1080/19345747.2022.2093299

*Cleary, T. J., Velardi, B., & Schnaidman, B. (2017). Effects of the self-regulation empowerment program (SREP) on middle school students' strategic skills, self-efficacy, and mathematics achievement. *Journal of School Psychology*, *64*, 28–42. https://doi.org/10.1016/j.jsp.2017.04.004

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal*, *50*(4), 812–850. https://doi.org/10.3102/0002831212469270

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, *97*(2), 31–47. https://doi.org/10.1257/aer.97.2.31

Currie, J., & Thomas, D. (2000). School quality and the longer-term effects of Head Start. *Journal of Human Resources*, *35*(4), 755–774. https://doi.org/10.2307/146372

Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K., & Sundararaman, V. (2013). School inputs, household substitution, and test scores. *American Economic Journal: Applied Economics*, *5*(2), 29–57. https://doi.org/10.1257/app.5.2.29

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics*, *1*(3), 111–134. https://doi.org/10.1257/app.1.3.111

Dietrichson, J., Bhatnagar, R., Filges, T., & Vembye, M. H. (2025). Replication material. *Open Science Framework*. https://osf.io/j6n5z/

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research*, *87*(2), 243–282. https://doi.org/10.3102/0034654316687036

Dietrichson, J., Filges, T., Klokker, R. H., Viinholt, B. C. A., Bøg, M., & Jensen, U. H. (2020). Targeted school-based interventions for improving reading and mathematics for students with, or at risk of, academic difficulties in Grades 7–12: A systematic review. *Campbell Systematic Reviews*, *16*(2), e1081. https://doi.org/10.1002/cl2.1081

Dietrichson, J., Filges, T., Seerup, J. K., Klokker, R. H., Viinholt, B. C. A., Bøg, M., & Eiberg, M. (2021). Targeted school-based interventions for improving reading and mathematics for students with or at risk of academic difficulties in Grades K-6: A systematic review. *Campbell Systematic Reviews*, *17*(2), e1152. https://doi.org/10.1002/cl2.1152

Dietrichson, J., Kristiansen, I. L., & Viinholt, B. A. (2020). Universal preschool programs and long-term child outcomes: A systematic review. *Journal of Economic Surveys*, *34*(5), 1007–1043. https://doi.org/10.1111/joes.12382

Dinerstein, M., Megalokonomou, R., & Yannelis, C. (2022). Human capital depreciation and returns to experience. *American Economic Review*, *112*(11), 3725–3762.

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology*, *32*, 271–297. https://doi.org/10.1146/annurev.soc.32.061604.123127

*Doabler, C. T., Clarke, B., Turtura, J. E., Sutherland, M., Gersib, J. A., Lesner, T., ... & Kosty, D. (2024). Proficiency with number concepts and operations: Replicating the efficacy of a first-grade mathematics intervention. *Journal of Learning Disabilities*, 57(4), 224–241. https://doi.org/10.1177/00222194231209017

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

Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, *27*(2), 109–132. https://doi.org/10.1257/jep.27.2.109

Durkin, K., Lipsey, M. W., Farran, D. C., & Wiesen, S. E. (2022). Effects of a statewide pre-kindergarten program on children's achievement and behavior through sixth grade. *Developmental Psychology*, *58*(3), 470–484. https://doi.org/10.1037/dev0001301

*Dyson, N. I., Jordan, N. C., & Glutting, J. (2013). A number sense intervention for low-income kindergartners at risk for mathematics difficulties. *Journal of Learning Disabilities*, *46*(2), 166–181. https://doi.org/10.1177/0022219411410233

*Dyson, N. I., Jordan, N. C., Beliakoff, A., & Hassinger-Das, B. (2015). A kindergarten number-sense intervention with contrasting practice conditions for low-achieving children. *Journal for Research in Mathematics Education*, *46*(3), 331–370. https://doi.org/10.5951/jresematheduc.46.3.0331

*Dyson, N. I., Jordan, N. C., Rodrigues, J., Barbieri, C., & Rinne, L. (2020). A fraction sense intervention for sixth graders with or at risk for mathematics difficulties. *Remedial and Special Education*, *41*(4), 244–254. https://doi.org/10.1177/0741932518807139

Egger, M., Juni, P., Bartlett, C., Holenstein, F., Sterne, J., et al. (2003). How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technology Assessment*, *7*(1), 1–76. https://doi.org/10.3310/hta7010

Eklund, K. M., Torppa, M., & Lyytinen, H. (2013). Predicting reading disability: Early cognitive risk and protective factors. *Dyslexia*, *19*(1), 1–10. https://doi.org/10.1002/dys.1447

*Elbro, C., & Petersen, D. K. (2004). Long-term effects of phoneme awareness and letter sound training: An intervention study with children at risk for dyslexia. *Journal of Educational Psychology*, *96*(4), 660–670. https://doi.org/10.1037/0022-0663.96.4.660

Fernandez-Castilla, B., Aloe, A. M., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2021). Estimating outcome-specific effects in meta-analyses of multiple outcomes: A simulation study. *Behavior Research Methods*, *53*(2), 702–717. https://doi.org/10.3758/s13428-020-01459-4

Filges, T., Sonne-Schmidt, C. S., & Nielsen, B. C. V. (2018). Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *Campbell Systematic Reviews*, *14*(1), 1–107. https://doi.org/10.4073/csr.2018.10

Fischer, M., Gerdtham, U.-G., Heckley, G., Karlsson, M., Kjellsson, G., & Nilsson, T. (2021). Education and health: Long-run effects of peers, tracking and years. *Economic Policy*, *36*(105), 3–49. https://doi.org/10.1093/epolic/eiaa027

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *Robumeta: Robust variance meta-regression*. Retrieved from https://github.com/zackfisher/robumeta

Fredriksson, P., Öckert, B., & Oosterbeek, H. (2013). Long-term effects of class size. *Quarterly Journal of Economics*, *128*(1), 249–285. https://doi.org/10.1093/qje/qjs048

Fredriksson, P., Öckert, B., & Tilley, J. L. (2024). *Parental and school responses to student performance: Evidence from school entry rules* (IZA Discussion Paper No. 16901). Retrieved from https://www.iza.org/publications/dp/16901/parental-and-school-responses-to-student-performance-evidence-from-school-entry-rules.

*Fuchs, L. S., Fuchs, D., Sterba, S. K., Barnes, M. A., Seethaler, P. M., & Changas, P. (2022). Building word-problem solving and working memory capacity: A randomized controlled trial comparing three intervention approaches. *Journal of Educational Psychology*, *114*(7), 1633–1653. https://doi.org/10.1037/edu0000752

*Fälth, L., Gustafson, S., Tjus, T., Heimann, M., & Svensson, I. (2013). Computer-assisted interventions targeting reading skills of children with reading disabilities–A longitudinal study. *Dyslexia*, *19*(1), 37–53. https://doi.org/10.1002/dys.1450

García, J. L., & Heckman, J. J. (2023). Parenting promotes social mobility within and across generations. *Annual Review of Economics*, *15*, 349–388. https://doi.org/10.1146/annurev-economics-021423-031905

Gathercole, S. E., Dunning, D. L., Holmes, J., & Norris, D. (2019). Working memory training involves learning new skills. *Journal of Memory and Language*, *105*, 19–42. https://doi.org/10.1016/j.jml.2018.10.003

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

Gelber, A., & Isen, A. (2013). Children's schooling and parents' behavior: Evidence from the Head Start Impact Study. *Journal of Public Economics*, *101*, 25–38. https://doi.org/10.1016/j.jpubeco.2013.02.005

Gensowski, M., Landersø, R., Dale, P., Bleses, Dorthe, Justice, L., & Hojen, A. (2020). *Public and parental investments and children's skill formation* (Study Paper No. 155). The Rockwool Foundation.

*Goodson, B., Wolf, A., Bell, S., Turner, H., & Finney, P. B. (2010). The effectiveness of a program to accelerate vocabulary development in kindergarten (VOCAB): Kindergarten final evaluation report. NCEE 2010-4014. *National Center for Education Evaluation and Regional Assistance*.

*Goodson, B., Wolf, A., Bell, S., Turner, H., & Finney, P. B. (2011). Effectiveness of a program to accelerate vocabulary development in kindergarten (VOCAB): First grade follow-up impact report and exploratory analyses of kindergarten impacts. Final report. NCEE 2012-4009. *National Center for Education Evaluation and Regional Assistance*.

Greaves, E., Hussain, I., Rabe, B., & Rasul, I. (2023). Parental responses to information about school quality: Evidence from linked survey and administrative data. *Economic Journal*, *133*(654), 2334–2402. https://doi.org/10.1093/ej/uead015

*Gunn, B., Smolkowski, K., Biglan, A., Black, C., & Blair, J. (2005). Fostering the development of reading skill through supplemental instruction: Results for Hispanic and non-Hispanic students. *Journal of Special Education*, *39*(2), 66–85. https://doi.org/10.1177/00224669050390020301

*Gustafson, S., Fälth, L., Svensson, I., Tjus, T., & Heimann, M. (2011). Effects of three interventions on the reading skills of children with reading disabilities in grade 2. *Journal of Learning Disabilities*, *44*(2), 123–135. https://doi.org/10.1177/0022219410391187

Halle, T. G., Hair, E. C., Buchinal, M., Anderson, R., & Zaslow, M. (2012). *In the running for successful outcomes: Exploring the evidence for thresholds of school readiness*. Technical report, U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation, Washington D. C. Retrieved from http://aspe.hhs.gov/reports/running-successful-outcomes-exploring-evidence-thresholds-school-readiness-technical-report-0.

*Hallstedt, M. H., Klingberg, T., & Ghaderi, A. (2018). Short and long-term effects of a mathematics tablet intervention for low performing second graders. *Journal of Educational Psychology*, *110*(8), 1127–1148. https://doi.org/10.1037/edu0000264

Hamman, E. A., Pappalardo, P., Bence, J. R., Peacor, S. D., & Osenberg, C. W. (2018). Bias in meta-analyses using Hedges' d. *Ecosphere*, *9*(9), e02419. https://doi.org/10.1002/ecs2.2419

*Hansson, Å. (2015). *Effekter av intensivundervisning i matematik. Utvärdering av ett pilotprojekt med personlig tränare i matematik för elever i behov av särskilt stöd* (RIPS: Rapporter Från Institutionen För Pedagogik Och Specialpedagogik No. nr 7). Institutionen för pedagogik och specialpedagogik.

Hart, E. R., Bailey, D. H., Luo, S., Sengupta, P., & Watts, T. W. (2024). Fadeout and persistence of intervention impacts on social-emotional and cognitive skills in children and adolescents: A meta-analytic review of randomized controlled trials. *Psychological Bulletin*, *150*(10), 1207–1236. https://doi.org/10.1037/bul0000450

*Hatz, H., & Sachse, S. (2010). Prävention von lese-rechtschreibstörungen: Auswirkungen eines trainings phonologischer bwusstheit und eines rechtschreibutrainings im ersten schuljahr auf den schriftspracherwerb bei risikokindern. *Zeitschrift für Entwicklungspsychologie Und Pädagogische Psychologie*, *42*(4), 226–240. https://doi.org/10.1026/0049-8637/a000021

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*(5782), 1900–1902. https://doi.org/10.1126/science.1128898

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, *103*(6), 2052–2086. https://doi.org/10.1257/aer.103.6.2052

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.2307/1164588

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Higgins, J. P. T., Eldridge, S., & Li, T. (2022). Chapter 23: Including variants on randomized trials. *Cochrane handbook for systematic reviews of interventions* (J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch, Eds.). Wiley Online Library. Retrieved from www.training.cochrane.org/handbook

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127–160. https://doi.org/10.1007/BF00401799

Hoover, W. A., & Tunmer, W. E. (2018). The simple view of reading: Three assessments of its adequacy. *Remedial and Special Education*, *39*(5), 304–312. https://doi.org/10.1177/0741932518773154

*Hurry, J., & Sylva, K. (2007). Long-term outcomes of early reading intervention. *Journal of Research in Reading*, *30*(3), 227–248. https://doi.org/10.1111/j.1467-9817.2007.00338.x

James, W. (1890). *Principles of psychology, volume one*. Dover Publications, Inc. https://doi.org/10.1037/10538-000

Jenkins, J. M., Watts, T. W., Magnuson, K., Gershoff, E. T., Clements, D. H., Sarama, J., & Duncan, G. J. (2018). Do high-quality kindergarten and first-grade classrooms mitigate preschool fadeout? *Journal of Research on Educational Effectiveness*, *11*(3), 339–374. https://doi.org/10.1080/19345747.2018.1441347

*Johnson, H., McNally, S., Rolfe, H., Ruiz-Valenzuela, J., Savage, R., Vousden, J., & Wood, C. (2019). Teaching assistants, computers and classroom management. *Labour Economics*, *58*, 21–36. https://doi.org/10.1016/j.labeco.2019.02.006

*Jones, C. J., & Christian, M. (2021). The results of a randomized control trial evaluation of the SPARK literacy program: An innovative approach that pairs one-on-one tutoring with

family engagement. *Journal of Education for Students Placed at Risk*, *26*(3), 185–209. https://doi.org/10.1080/10824669.2020.1809419

*Jones, C. J., Reeves, M., Li, D., & Gilman, L. (2023). What is the sustained impact of Future Forward on reading achievement, attendance, and special education placement 5 years after participation? *Educational Evaluation and Policy Analysis*, online advance access. https://doi.org/10.3102/016237372311826

*Jordan, N. C., Glutting, J., Dyson, N., Hassinger-Das, B., & Irwin, C. (2012). Building kindergartners' number sense: A randomized controlled study. *Journal of Educational Psychology*, *104*(3), 647–660. https://doi.org/10.1037/a0029018

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). Arithmetic fact mastery in young children: A longitudinal investigation. *Journal of Experimental Child Psychology*, *85*(2), 103–119. https://doi.org/10.1016/S0022-0965(03)00032-8

Kang, C. Y., Duncan, G. J., Clements, D. H., Sarama, J., & Bailey, D. H. (2019). The roles of transfer of learning and forgetting in the persistence and fadeout of early childhood mathematics interventions. *Journal of Educational Psychology*, *111*(4), 590–603. https://doi.org/10.1037/edu0000297

Kitchens, K. E., Gormley, W., & Anderson, S. (2020). Do better schools help to prolong early childhood education effects? *Journal of Applied Developmental Psychology*, *66*, 101092. https://doi.org/10.1016/j.appdev.2019.101092

Knudsen, E. I., Heckman, J. J., Cameron, J. L., & Shonkoff, J. P. (2006). Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proceedings of the National Academy of Sciences*, *103*(27), 10155–10162. https://doi.org/10.1073/pnas.0600888103

*Konerding, M., Bergström, K., Lachmann, T., & Klatte, M. (2020). Effects of computerized grapho-phonological training on literacy acquisition and vocabulary knowledge in children with an immigrant background learning German as L2. *Journal of Cultural Cognitive Science*, *4*, 368–383. https://doi.org/10.1007/s41809-020-00064-3

Li, W., Duncan, G. J., Magnuson, K., Schindler, H. S., Yoshikawa, H., & Leak, J. (2020). *Timing in early childhood education: How cognitive and achievement program impacts vary by starting age, program duration, and time since the end of the program* (EdWorkingPaper: 20-201). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/5tvg-nt21

Lin, L., & Aloe, A. M. (2021). Evaluation of various estimators for standardized mean difference in meta-analysis. *Statistics in Medicine*, *40*(2), 403–426. https://doi.org/10.1002/sim.8781

Linden, A. (2013). Assessing regression to the mean effects in health care initiatives. *BMC Medical Research Methodology*, *13*(1), 1–7. https://doi.org/10.1186/1471-2288-13-119

*Lindström-Sandahl, H., Danielsson, H., Elwér, Å., Samuelsson, S., & Samuelsson, J. (2024). A 1-year follow-up of two intensive supplemental reading and mathematics programs in Swedish elementary school. In Lindström-Sandahl, H. *Early elementary school interventions in reading and mathematics* (Doctoral dissertation, Linköping University Electronic Press). Linköping University. https://doi.org/10.3384/9789180754750

Lipsey, M. W., Farran, D. C., & Durkin, K. (2018). Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly*, *45*, 155–176. https://doi.org/10.1016/j.ecresq.2018.03.005

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *National Center for Special Education Research*.

Logan, G. D. (1997). Automaticity and reading: Perspectives from the instance theory of automatization. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *13*(2), 123–146. https://doi.org/10.1080/1057356970130203

*Lopez-Pedersen, A., Mononen, R., Aunio, P., Scherer, R., & Melby-Lervåg, M. (2023). Improving numeracy skills in first graders with low performance in early numeracy: A randomized controlled trial. *Remedial and Special Education*, *44*(2), 126–136. https://doi.org/10.1177/07419325221102537

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25. https://doi.org/10.1086/392759

Mantzicopoulos, P., Morrison, D., Stone, E., & Setrakian, W. (1992). Use of the SEARCH/TEACH tutoring approach with middle-class students at risk for reading failure. *The Elementary School Journal*, *92*(5), 573–586. https://doi.org/10.1086/461707

Marin-Martinez, F., & Sanchez-Meca, J. (2010). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement*, *70*(1), 56–73. https://doi.org/10.1177/0013164409344534

Martin, B. N. (2021). *Fadeout or forgetting? A brief booster lesson to support the maintenance of first-grade mathematics intervention effects* (Doctoral dissertation, Vanderbilt University). Graduate School of Vanderbilt University.

Masten, A. S., & Cicchetti, D. (2010). Developmental cascades. *Development and Psychopathology*, *22*(3), 491–495. https://doi.org/10.1017/S0954579410000222

McCormick, M., Weiland, C., Hsueh, J., Pralica, M., Weissman, A. K., Moffett, L., ... Sachs, J. (2021). Is skill type the key to the PreK fadeout puzzle? Differential associations between enrollment in PreK and constrained and unconstrained skills across kindergarten. *Child Development*, *92*(4), e599–e620. https://doi.org/10.1111/cdev.13520

McCoy, D. C., Yoshikawa, H., Ziol-Guest, K. M., Duncan, G. J., Schindler, H. S., Magnuson, K., ... Shonkoff, J. P. (2017). Impacts of early childhood education on medium-and long-term educational outcomes. *Educational Researcher*, *46*(8), 474–487. https://doi.org/10.3102/0013189X17737739

Meghir, C., & Palme, M. (2005). Educational reform, ability, and family background. *American Economic Review*, *95*(1), 414–424. https://doi.org/10.1257/0002828053828671

Montessori, M. (1912). *The Montessori method*. Frederick A. Stokes Company.

National Mathematics Advisory Panel (2008). *Foundations for success: The final report of the national mathematics advisory panel*. US Department of Education.

Nelson, P. M., Klingbeil, D. A., Van Norman, E. R., & Parker, D. C. (2021). A cluster randomized controlled trial of brief follow-up practice sessions on intervention maintenance. *Journal of School Psychology*, *88*, 31–46. https://doi.org/10.1016/j.jsp.2021.07.003

Nickow, A., Oreopoulos, P., & Quan, V. (2024). The promise of tutoring for PreK–12 learning: A systematic review and meta-analysis of the experimental evidence. *American Educational Research Journal*, *61*(1), 74–107. https://doi.org/10.3102/000283122312086

Pace, A., Alper, R., Burchinal, M. R., Golinkoff, R. M., & Hirsh-Pasek, K. (2019). Measuring success: Within and cross-domain predictors of academic and social trajectories in elementary school. *Early Childhood Research Quarterly*, *46*, 112–125. https://doi.org/10.1016/j.ecresq.2018.04.001

Pages, R., Bailey, D. H., & Duncan, G. J. (2023). The impacts of Abecedarian and Head Start on educational attainment: Reasoning about unobserved mechanisms from temporal patterns of indirect effects. *Early Childhood Research Quarterly*, 65, 261–274. https://doi.org/10.1016/j.ecresq.2023.07.003

Pfost, M., Hattie, J., Dörfler, T., & Artelt, C. (2014). Individual differences in reading development: A review of 25 years of empirical research on Matthew effects in reading. *Review of Educational Research*, *84*(2), 203–244. https://doi.org/10.3102/0034654313509492

Pop-Eleches, C., & Urquiola, M. (2013). Going to a better school: Effects and behavioral responses. *American Economic Review*, *103*(4), 1289–1324. https://doi.org/10.1257/aer.103.4.1289

*Powell, S. R., Berry, K. A., Fall, A.-M., Roberts, G., Fuchs, L. S., & Barnes, M. A. (2021). Alternative paths to improved word-problem performance: An advantage for embedding prealgebraic reasoning instruction within word-problem intervention. *Journal of Educational Psychology*, *133*(5), 898–910. https://doi.org.10.1037/edu0000513

*Powell, S. R., Berry, K. A., Fall, A. M., Roberts, G., Barnes, M. A., Fuchs, L. S., ... & Lin, X. (2022). Does word-problem performance maintain? Follow-up one year after implementation of a word-problem intervention. *Journal of Research on Educational Effectiveness*, *15*(1), 52–77. https://doi.org/10.1080/19345747.2021.1961332

*Prediger, S., & Wessel, L. (2018). Brauchen mehrsprachige jugendliche eine andere fach-und sprachintegrierte förderung als einsprachige? *Zeitschrift für Erziehungswissenschaft*, *21*(2), 361–382. https://doi.org/10.1007/s11618-017-0785-8

Protopapas, A., Parrila, R., & Simos, P. G. (2016). In search of Matthew effects in reading. *Journal of Learning Disabilities*, *49*(5), 499–514. https://doi.org/10.1177/0022219414559974

Protzko, J. (2015). The environment in raising early intelligence: A meta-analysis of the fadeout effect. *Intelligence*, *53*, 202–210. https://doi.org/10.1016/j.intell.2015.10.006

Protzko, J. (2016). Does the raising IQ-raising g distinction explain the fadeout effect? *Intelligence*, *56*, 65–71. https://doi.org/10.1016/j.intell.2016.02.008

Pustejovsky, J. (2022). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections*. Retrieved from http://jepusto.github.io/clubSandwich/

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, *10*(1), 57–71. https://doi.org/10.1002/jrsm.1332

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*, 425–438. https://doi.org/10.1007/s11121-021-01246-3

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, *52*(1), 501–525. https://doi.org/10.1146/annurev.psych.52.1.501

Raver, C. C. (2012). Low-income children's self-regulation in the classroom: Scientific inquiry for social change. *American Psychologist*, *67*(8), 681–689. https://doi.org/10.1037/a0030085

Reynolds, A. J., & Ou, S. R. (2011). Paths of effects from preschool to adult well-being: A confirmatory analysis of the child-parent center program. *Child Development*, *82*(2), 555–582. https://doi.org/10.1111/j.1467-8624.2010.01562.x

Rosengarten, M. L., Hart, E. R., Bailey, D. H., McCormick, M. P., Lovett, B. J., & Watts, T. W. (2024). *Using meta-analytic data to examine fadeout and persistence of intervention impacts on constrained and unconstrained skills* (EdWorkingPaper No. 24-1069). Retrieved from Annenberg Institute for School Reform at Brown University: https://doi.org/10.26300/qn5q-ms33

*Ryder, J. F., Tunmer, W. E., & Greaney, K. T. (2008). Explicit instruction in phonemic awareness and phonemically based decoding skills as an intervention strategy for struggling readers in whole language classrooms. *Reading and Writing*, *21*(4), 349–369. https://doi.org/10.1007/s11145-007-9080-z

Sánchez-Meca, J., & Marin-Martinez, F. (1998). Weighting by inverse variance or by sample size in meta-analysis: A simulation study. *Educational and Psychological Measurement*, *58*(2), 211–220. https://doi.org/10.1177/0013164498058002005

Savi, A. O., Marsman, M., van der Maas, H. L., & Maris, G. K. (2019). The wiring of intelligence. *Perspectives on Psychological Science*, *14*(6), 1034–1061. https://doi.org/10.1177/1745691619866447

Scammacca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A meta-analysis of interventions for struggling readers in grades 4–12: 1980–2011. *Journal of Learning Disabilities*, *48*(4), 369–390. https://doi.org/10.1177/0022219413504995

*Schanzenbach, D. W. (2006). What have researchers learned from Project STAR? *Brookings Papers on Education Policy*, *9*, 205–228. https://doi.org/http://www.jstor.org/stable/20067282

Schneider, B., & Bradford, L. (2020). What we are learning about fade-out of intervention effects: A commentary. *Psychological Science in the Public Interest*, *21*(2), 50–54. https://doi.org/10.1177/1529100620935793

Schroeders, U., Schipolowski, S., Zettler, I., Golle, J., & Wilhelm, O. (2016). Do the smart get smarter? Development of fluid and crystallized intelligence in 3rd grade. *Intelligence*, *59*, 84–95. https://doi.org/10.1016/j.intell.2016.08.003

*Schüler-Meyer, A., Prediger, S., Kuzu, T., Wessel, L., & Redder, A. (2019). Is formal language proficiency in the home language required to profit from a bilingual teaching intervention in mathematics? A mixed methods study on fostering multilingual students' conceptual understanding. *International Journal of Science and Mathematics Education*, *17*(2), 317–339. https://doi.org/10.1007/s10763-017-9857-8

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*(2), 151–218. https://doi.org/10.1016/0010-0277(94)00645-2

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453.

Slavin, R. E. (1995). Best evidence synthesis: An intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology*, *48*(1), 9–18. https://doi.org/10.1016/0895-4356(94)00097-a

Smithers, L. G., Sawyer, A. C. P., Chittleborough, C. R., Davies, N. M., Davey Smith, G., & Lynch, J. W. (2018). A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes. *Nature Human Behaviour*, *2*(11), 867–880. https://doi.org/10.1038/s41562-018-0461-x

*Somers, M.-A., Corrin, W., Sepanik, S., Salinger, T., Levin, J., & Zmach, C. (2010). *The enhanced reading opportunities study. Final report: The impact of supplemental literacy*

courses for struggling ninth-grade readers (NCEE 2010-4021). National Center for Education Evaluation and Regional Assistance.

Sood, S., & Jitendra, A. K. (2013). An exploratory study of a number sense program to develop kindergarten students' number proficiency. *Journal of Learning Disabilities*, *46*(4), 328–346. https://doi.org /10.1177/0022219411422380

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407. https://doi.org/10.1177/0022057409189001-204

Student, S. R. (2022). Vertical scales, deceleration, and empirical benchmarks for growth. *Educational Researcher*, *51*(8), 536–543. https://doi.org/10.3102/0013189X221105873

Suggate, S. P. (2016). A meta-analysis of the long-term effects of phonemic awareness, phonics, fluency, and reading comprehension interventions. *Journal of Learning Disabilities*, *49*(1), 77–96. https://doi.org/10.1177/0022219414528540

*Svensson, I., Nordström, T., Lindeblad, E., Gustafson, S., Björn, M., Sand, C., ... & Nilsson, S. (2021). Effects of assistive technology for students with reading and writing disabilities. *Disability and Rehabilitation: Assistive Technology*, *16*(2), 196-208. https://doi.org/10.1080/17483107.2019.1646821

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, *120*(3), 439–471. https://doi.org/10.1037/a0033138

Teicher, M. H., Samson, J. A., Anderson, C. M., & Ohashi, K. (2016). The effects of childhood maltreatment on brain structure, function and connectivity. *Nature Reviews Neuroscience*, *17*(10), 652–666. https://doi.org/10.1038/nrn.2016.111

The Methods Group of the Campbell Collaboration (2019). *Methodological expectations of Campbell Collaboration intervention reviews: Reporting standards*. Campbell Policies and Guidelines Series. https://doi.org/10.4073/cpg.2016.4

Thorndike, E., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions (i). *Psychological Review*, *8*(3), 247–261. https://doi.org/10.1037/h0074898

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375–393. https://doi.org/10.1037/met0000011

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*(6), 604–634. https://doi.org/10.3102/1076998615606099

*Tolan, P., Gorman-Smith, D., & Henry, D. (2004). Supporting families in a high-risk setting: Proximal effects of the SAFEChildren preventive intervention. *Journal of Consulting and Clinical Psychology*, *72*(5), 855–869. https://doi.org/10.1037/0022-006X.72.5.855

*Toll, S. W., & Van Luit, J. E. (2014). Effects of remedial numeracy instruction throughout kindergarten starting at different ages: Evidence from a large-scale longitudinal study. *Learning and Instruction*, *33*, 39–49. https://doi.org/10.1016/j.learninstruc.2014.03.003

*Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Herron, J., & Lindamood, P. (2010). Computer-assisted instruction to prevent early reading difficulties in students at risk for dyslexia: Outcomes from two instructional approaches. *Annals of Dyslexia*, *60*(1), 40–56. https://doi.org/10.1007/s11881-009-0032-y

*Torgesen, J., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., ... Haan, C. (2006). National assessment of Title I: Interim report. Volume II: Closing the reading gap:

First year findings from a randomized trial of four reading interventions for striving readers. NCEE 2006-4002. *National Center for Education Evaluation and Regional Assistance*.

*Torgesen, J., Schirm, A., Castner, L., Vartivarian, S., Mansfield, W., Myers, D., ... Haan, C. (2007). National assessment of Title I. Final report. Volume II: Closing the reading gap–findings from a randomized trial of four reading interventions for striving readers. NCEE 2008-4013. *National Center for Education Evaluation and Regional Assistance*.

Toste, J. R., Didion, L., Peng, P., Filderman, M. J., & McClelland, A. M. (2020). A meta-analytic review of the relations between motivation and reading achievement for K–12 students. *Review of Educational Research*, *90*(3), 420–456. https://doi.org/10.3102/0034654320919352

*Vadasy, P. F., Jenkins, J. R., & Pool, K. (2000). Effects of tutoring in phonological and early reading skills on students at risk for reading disabilities. *Journal of Learning Disabilities*, *33*(6), 579–590. https://doi.org/10.1177/002221940003300606

*Vadasy, P. F., & Sanders, E. A. (2010). Efficacy of supplemental phonics-based instruction for low-skilled kindergarteners in the context of language minority status and classroom phonics instruction. *Journal of Educational Psychology*, *102*(4), 786–803. https://doi.org/10.1037/a0019639

*Vadasy, P. F., & Sanders, E. A. (2011). Efficacy of supplemental phonics-based instruction for low-skilled first graders: How language minority status and pretest characteristics moderate treatment response. *Scientific Studies of Reading*, *15*(6), 471–497. https://doi.org/10.1080/10888438.2010.501091

*Vadasy, P. F., & Sanders, E. A. (2012). Two-year follow-up of a kindergarten phonics intervention for English learners and native English speakers: Contextualizing treatment impacts by classroom literacy instruction. *Journal of Educational Psychology*, *104*(4), 987–1005. https://doi.org/10.1037/a0028163

*Vadasy, P. F., & Sanders, E. A. (2013). Two-year follow-up of a code-oriented intervention for lower-skilled first-graders: The influence of language status and word reading skills on third-grade literacy outcomes. *Reading and Writing*, *26*(6), 821–843. https://doi.org/10.1007/s11145-012-9393-4

*Van De Rijt, B. A., & Van Luit, J. E. (1998). Effectiveness of the additional early mathematics program for teaching children early mathematics. *Instructional Science*, *26*(5), 337–358. https://doi.org/10.1023/A:1003180411209

Van den Noortgate, W., Lopez-Lopez, J. A., Marin-Martinez, F., & Sanchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

Van den Noortgate, W., Lopez-Lopez, J. A., Marin-Martinez, F., & Sanchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods*, *47*(4), 1274–1294. https://doi.org/10.3758/s13428-014-0527-2

Van Der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842

*Vaughn, S., Cirino, P. T., Linan-Thompson, S., Mathes, P. G., Carlson, C. D., Hagan, E. C., ... Francis, D. J. (2006). Effectiveness of a Spanish intervention and an English intervention

for English-language learners at risk for reading problems. *American Educational Research Journal*, *43*(3), 449–487. https://doi.org/10.3102/00028312043003449

*Vaughn, S., Cirino, P. T., Tolar, T., Fletcher, J. M., Cardenas-Hagan, E., Carlson, C. D., & Francis, D. J. (2008). Long-term follow-up of Spanish and English interventions for first-grade English language learners at risk for reading problems. *Journal of Research on Educational Effectiveness*, *1*(3), 179–214. https://doi.org/10.1080/19345740802114749

*Vaughn, S., Linan-Thompson, S., Mathes, P. G., Cirino, P. T., Carlson, C. D., Pollard-Durodola, S. D., ... Francis, D. J. (2006). Effectiveness of Spanish intervention for first-grade English language learners at risk for reading difficulties. *Journal of Learning Disabilities*, *39*(1), 56–73. https://doi.org/10.1177/00222194060390010601

*Vaughn, S., Mathes, P., Linan-Thompson, S., Cirino, P., Carlson, C., Pollard-Durodola, S., ... Francis, D. (2006). Effectiveness of an English intervention for first-grade English language learners at risk for reading problems. *The Elementary School Journal*, *107*(2), 153–180. https://doi.org/10.1086/510653

Vembye, M. H., Pustejovsky, J. E., & Pigott, T. D. (2023). Power approximations for overall average effects in meta-analysis with dependent effect sizes. *Journal of Educational and Behavioral Statistics*, *48*(1), 70–102. https://doi.org/10769986221127379

Viechtbauer, W. (2022). *Metafor: Meta-analysis package for r*. Retrieved from https://CRAN.R-project.org/package=metafor.

Vlach, H. A. (2019). Learning to remember words: Memory constraints as double-edged sword mechanisms of language development. *Child Development Perspectives*, *13*(3), 159–165. https://doi.org/10.1111/cdep.12337

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in reading and math skills forms mainly before kindergarten: A replication, and partial correction, of "Are schools the great equalizer?". *Sociology of Education*, *91*(4), 323–357. https://doi.org/10.1177/0038040718801760

Vygotsky, L. S. (2012). *Thought and language*. Cambridge, MA: MIT press.

Wan, S., Bond, T. N., Lang, K., Clements, D. H., Sarama, J., & Bailey, D. H. (2021). Is intervention fadeout a scaling artefact? *Economics of Education Review*, *82*, 102090. https://doi.org/10.1016/j.econedurev.2021.102090

Wasik, B. A., & Slavin, R. E. (1993). Preventing early reading failure with one-to-one tutoring: A review of five programs. *Reading Research Quarterly*, 179–200. https://doi.org/10.2307/747888

Watts, T. W., Bailey, D. H., & Li, C. (2019). Aiming further: Addressing the need for high-quality longitudinal research in education. *Journal of Research on Educational Effectiveness*, *12*(4), 648–658. https://doi.org/10.1080/19345747.2019.1644692

Watts, T. W., Clements, D. H., Sarama, J., Wolfe, C. B., Spitler, M. E., & Bailey, D. H. (2017). Does early mathematics intervention change the processes underlying children's learning? *Journal of Research on Educational Effectiveness*, *10*(1), 96–115. https://doi.org/10.1080/19345747.2016.1204640

Watts, T. W., Duncan, G. J., & Rivas, M. (2019). *A reanalysis of impacts of the Tennessee Voluntary Prekindergarten Program* (EdWorkingPaper No. 19-28). Retrieved from Annenberg Institute at Brown University: http://edworkingpapers.com/ai19-28.

Watts, T., Ibrahim, D., Khader, A., Li, C., Gandhi, J., & Raver, C. (2020). Exploring the impacts of an early childhood educational intervention on later school selection. *Educational Researcher*, *49*(9), 667–677. https://doi.org/10.3102/0013189X20935060

What Works Clearinghouse. (2021). *Supplement document for appendix e and the What Works Clearinghouse procedures handbook, version 4.1.* U.S. Department of Education, Institute for Education Sciences, National Center For Education Evaluation; Regional Assistance (NCEE).

Whitehurst, G. J. (2018). Commentary: The positive impacts of public pre-k fade quickly, and sometimes reverse: What does this portend for future research and policy? *Early Childhood Research Quarterly*, *45*, 183–187. https://doi.org/10.1016/j.ecresq.2018.02.016

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*, 235–269. https://doi.org/10.1146/annurev.psych.55.090902.141555

Wolf, B., & Harbatkin, E. (2022). Making sense of effect sizes: Systematic differences in intervention effect sizes by outcome measure type. *Journal of Research on Educational Effectiveness*, 1–28. https://doi.org/10.1080/19345747.2022.2071364

Wolf, S., Aber, J. L., Behrman, J. R., & Peele, M. (2019). Longitudinal causal impacts of preschool teacher training on Ghanaian children's school readiness: Evidence for persistence and fade-out. *Developmental Science*, *22*(5), e12878. https://doi.org/10.1111/desc.12878

*Wolff, U. (2011). Effects of a randomised reading intervention study: An application of structural equation modelling. *Dyslexia*, *17*(4), 295–311. https://doi.org/10.1002/dys.438

*Wolff, U. (2016). Effects of a randomized reading intervention study aimed at 9-year-olds: A 5-year follow-up. *Dyslexia*, *22*(2), 85–100. https://doi.org/10.1002/dys.1529

Supplemental Material to "On the Mechanisms of Intervention Effect Fade-Out: A Meta-Analytic Review of Interventions Targeting At-Risk Students' Achievement"

Table of Contents

A: Sample Description S2

B: Identification of Identical Test Items and Sequential Test of H5a S9

C: Blinding of Testers S15

D: Full Results of the Sensitivity Analyses S16

E: Assumptions and Calculations of the Regression to the Mean Effect S44

F: Test of Assumption for Increasing Variance Mechanism S51

G: Assumption on Trends in Relation to Outside Sample S54

References S60

Supplemental Material A: Sample Description

Table S1 describes how the 53 articles and reports (records) are clustered in 42 studies. The study numbers in the table correspond to the numbers in Figure 1. Columns 3-6 contain, calculated separately for post and follow-up, the study-average effect size and study-average sample size in terms of students. The study-average effect size is the mean of the cluster-adjusted *g*'s by study, and the study-average sample size is the mean of the sample sizes across *g*'s by study.[1] The study-average sample size is usually smaller at follow-up due to attrition but is in some studies larger, e.g., because the studies reported results for combined treatments, in which case the sample size for each effect size increased. The rightmost column indicates in which review the records were included, or if the record was found through forward citation-tracking.

Table S2 shows study-level averages of the outcomes measures and the moderators used in our primary analysis. For our full coding of each record and the extracted data, see the supplementary material at https://osf.io/j6n5z/.

[1] Note that the average effect size at post-test in this table does not correspond exactly to the average in Figure 1 for all studies. The post-test effect sizes in Figure 1 excluded a few outliers whereas the averages in Table S1 are calculated across all post-test effect sizes (this difference is the reason why the ranking of studies by average effect size is not the same in Table S1 as in Figure 1).

Table S1

Records and studies in full sample

(1) Study id	(2) Records	(3) ES, post	(4) Sample size, post	(5) ES, follow-up	(6) Sample size, follow-up	(7) Included in review/citation-tracking
1	Ryder et al. (2008)	1.78	24	0.79	20	Author (2021a)
2	Van De Rijt & Van Luit (1998)	1.12	50	0.79	50	Author (2021a)
3	Fuchs et al. (2022)	1.07	118	0.72	118	Citation-tracking
4	Toll & Van Luit (2014)	0.97	256	0.41	242	Author (2021a)
5	Vadasy et al. (2000)	0.95	46	0.65	37	Author (2017, 2021a)
6	Dyson et al. (2020)	0.95	52	0.82	52	Citation-tracking
7	Blachman et al. (2004, 2014)	0.90	69	0.41	63	Author (2021a)
8	Powell et al. (2021, 2022)	0.77	160	0.13	160	Citation-tracking
9	Prediger & Wessel (2018)	0.64	250	0.32	250	Author (2020b)
10	Dyson et al. (2015)	0.57	84	0.36	84	Author (2021a)
11	Vadasy & Sanders (2010, 2012)	0.57	74	0.38	66	Author (2017, 2021a)
12	Jordan et al. (2012)	0.55	86	0.32	86	Author (2017, 2021a)
13	Torgesen et al. (2010)	0.53	76	0.36	74	Author (2021a)
14	Schüler-Meyer et al. (2019)	0.52	86	0.10	86	Author (2020b)
15	Lindström-Sandahl et al. (2024)	0.42	64	0.33	60	Citation-tracking
16	Bailey et al. (2020)	0.38	410	0.06	366	Citation-tracking
17	Vaughn et al. (2006a, 2006b, 2006c); Vaughn et al. (2008); Cirino et al. (2009)	0.40	67	0.31	98	Author (2021a)
18	Hansson (2015)	0.37	57	0.17	57	Author (2017, 2021a)
19	Hallstedt et al. (2018)	0.35	128	0.17	128	Author (2021a)
20	Gunn et al. (2005)	0.33	114	0.22	97	Author (2021a)
21	Konerding et al. (2020)	0.25	26	0.12	25	Citation-tracking
22	Clarke et al. (2016)	0.25	259	-0.02	197	Author (2021a)
23	Johnson et al. (2019)	0.23	1402	0.09	1318	Citation-tracking
24	Vadasy & Sanders (2011, 2013)	0.23	94	0.20	88	Author (2017, 2021a)
25	Alan et al. (2019)	0.22	1348	0.12	780	Citation-tracking
26	Elbro & Klint (2004)	0.18	82	0.29	82	Author (2021a)
27	Doabler et al. (2024)	0.20	140	0.23	128	Citation-tracking
28	Tolan et al. (2004)	0.17	363	0.17	369	Author (2017, 2021a)
29	Schanzenbach (2006)	0.15	4259	0.04	3034	Author (2018)
30	Clarke et al. (2023)	0.15	263	0.05	225	Citation-tracking

(1) Study id	(2) Records	(3) ES, post	(4) Sample size, post	(5) ES, follow-up	(6) Sample size, follow-up	(7) Included in review/citation-tracking
31	Torgesen et al. (2006, 2007)	0.14	93	0.12	91	Author (2017, 2021a)
32	Goodson et al. (2010, 2011)	0.13	1296	0.03	1296	Author (2017)
33	Dyson et al. (2013)	0.11	121	0.00	121	Author (2017, 2021a)
34	Gustafson et al. (2011); Fälth et al. (2013)	0.10	50	0.30	50	Author (2021a)
35	Jones & Christian (2021); Jones et al. (2023)	0.10	378	0.16	355	Citation-tracking
36	Somers et al. (2010)	0.09	1228	-0.01	1236	Author (2020b)
37	Wolff (2011, 2016)	0.08	112	0.12	98	Author (2021a)
38	Hatz & Sachse (2010)	0.08	81	-0.11	70	Author (2021a)
39	Lopez-Pedersen et al. (2023)	0.04	118	0.02	113	Citation-tracking
40	Svensson et al. (2021)	-0.05	63	-0.12	51	Citation-tracking
41	Hurry & Sylva (2007)	-0.12	131	-0.14	125	Author (2021a)
42	Cleary et al. (2017)	-0.34	37	0.44	37	Author (2020b)

Note. The columns contain: (1) the study identification number, which corresponds to the number in Figure 1; (2) the records included in the study; (3) the post-test study-average effect size; (4) the post-test study-average sample size; (5) the follow-up study-average effect size; (6) the follow-up study-average sample size; and (7) the review or reviews in which the records were included, or if the record(s) was found through forward

Table S2

Outcome measures and moderators

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Study	Absolute fade-out	ES, raw scores on identical tests	Absolute fade-out, identical tests	ES, outside sample - T	ES, outside sample - C	Identical samples	Duration (12-17 months)	Duration (>17 months)	Post-test ES (anchors)	Identical tests
Alan et al. (2019)	-0.101	NA	NA	NA	NA	0.000	0.000	1.000	0.217	0.000
Bailey et al. (2020)	-0.168	NA	NA	-0.264	-0.372	0.000	0.500	0.500	0.224	0.000
Blachman et al. (2004, 2014)	-0.506	NA	NA	-0.836	-1.199	0.467	0.467	0.533	0.915	0.467
Clarke et al. (2016a)	-0.201	NA	NA	NA	NA	0.000	0.000	0.000	0.179	0.000
Clarke et al. (2023)	-0.068	NA	NA	NA	NA	0.000	0.000	0.000	0.117	0.000
Cleary et al. (2017)	0.777	NA	NA	-0.002	-0.262	1.000	0.250	0.000	-0.339	0.000
Doabler et al. (2023)	0.105	NA	NA	NA	NA	0.000	0.000	0.000	0.123	0.000
Dyson et al. (2013)	-0.104	0.116	-0.104	-0.199	-0.293	1.000	0.000	0.000	0.105	1.000
Dyson et al. (2015)	-0.211	0.316	-0.211	-1.069	-1.945	1.000	0.000	0.000	0.568	1.000
Dyson et al. (2020)	-0.120	0.141	-0.120	NA	NA	0.000	0.000	0.000	0.945	1.000
Elbro & Klint (2004)	-0.072	NA	NA	-0.280	-0.508	0.000	0.250	0.750	0.359	0.000
Fuchs et al. (2022)	-0.346	NA	NA	NA	NA	1.000	0.000	0.000	1.070	0.000
Goodson et al. (2012)	-0.102	NA	NA	-0.564	-0.618	0.000	1.000	0.000	0.133	0.000
Gunn et al. (2005)	-0.105	NA	NA	-0.526	-0.865	0.000	0.000	1.000	0.326	0.000
Gustafson et al. (2011); Fälth et al. (2013)	0.200	0.624	0.200	-1.759	-1.964	1.000	0.000	0.000	0.103	1.000
Hallstedt et al. (2018)	-0.175	0.436	-0.175	NA	NA	0.000	0.500	0.000	0.348	1.000
Hansson (2014)	-0.195	-0.204	-0.195	-0.426	-0.452	1.000	0.000	0.000	0.367	1.000
Hatz & Sachse (2010)	-0.233	1.656	-0.185	NA	NA	0.000	1.000	0.000	0.121	0.500

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Study	Absolute fade-out	ES, raw scores on identical tests	Absolute fade-out, identical tests	ES, outside sample - T	ES, outside sample - C	Identical samples	Duration (12-17 months)	Duration (>17 months)	Post-test ES (anchors)	Identical tests
Hurry & Sylva (2007)	-0.027	0.803	-0.008	NA	NA	0.000	0.500	0.500	-0.115	0.500
Johnson et al. (2019)	-0.145	NA	NA	NA	NA	0.000	1.000	0.000	0.234	0.000
Jones & Christian (2021); Jones et al. (2023)	0.060	NA	NA	NA	NA	0.000	0.200	0.800	0.100	0.000
Jordan et al. (2012)	-0.226	0.106	-0.226	-0.157	-0.391	1.000	0.000	0.000	0.547	1.000
Konerding et al. (2020)	-0.129	0.268	-0.129	NA	NA	0.000	0.000	0.000	0.250	1.000
Lindström-Sandahl et al. (2024)	-0.092	0.790	-0.092	-0.199	-0.551	0.000	1.000	0.000	0.423	1.000
Lopez-Pedersen et al. (2023)	-0.024	0.474	-0.024	NA	NA	0.000	0.000	0.000	0.044	1.000
Powell et al. (2022)	-0.634	-0.162	-0.634	NA	NA	1.000	0.000	0.000	0.767	1.000
Prediger & Wessel (2018)	-0.328	NA	NA	NA	NA	1.000	0.000	0.000	0.643	0.000
Ryder et al. (2008)	-0.026	1.337	-0.026	NA	NA	0.000	0.000	1.000	0.819	1.000
Schanzenbach (2007)	-0.108	NA	NA	NA	NA	0.000	0.200	0.800	0.152	0.000
Schüler-Meyer et al. (2019)	-0.420	-0.049	-0.420	NA	NA	1.000	0.000	0.000	0.517	1.000
Somers et al. (2010)	-0.105	NA	NA	NA	NA	0.000	1.000	0.000	0.090	0.000
Svensson et al. (2021)	-0.075	0.244	-0.075	NA	NA	0.000	1.000	0.000	-0.045	1.000
Tolan et al. (2004)	0.000	NA	NA	NA	NA	0.000	0.000	0.000	0.174	0.000
Toll et al. (2014)	-0.556	NA	NA	-0.621	-1.339	0.000	0.000	0.000	0.965	0.000
Torgesen et al. (2006, 2007)	-0.026	NA	NA	-0.545	-0.710	0.000	1.000	0.000	0.142	0.000
Torgesen et al. (2010)	-0.148	0.823	-0.137	0.172	-0.229	0.000	1.000	0.000	0.505	0.769

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Study	Absolute fade-out	ES, raw scores on identical tests	Absolute fade-out, identical tests	ES, outside sample - T	ES, outside sample - C	Identical samples	Duration (12-17 months)	Duration (>17 months)	Post-test ES (anchors)	Identical tests
Vadasy & Sanders (2010, 2012)	-0.315	NA	NA	-0.217	-0.506	0.000	0.250	0.500	0.694	0.000
Vadasy & Sanders (2011, 2013)	-0.091	NA	NA	-0.056	-0.268	0.000	0.500	0.250	0.289	0.000
Vadasy et al. (2000)	-0.274	NA	NA	-0.110	-0.765	0.000	1.000	0.000	0.920	0.250
Van de Rijt et al. (1998)	-0.334	NA	NA	NA	NA	1.000	0.000	0.000	1.122	1.000
Vaughn et al. (2006a, 2006b, 2006c); Vaughn et al. (2008); Cirino et al. (2009)	-0.090	NA	NA	-0.489	-0.874	0.000	0.444	0.556	0.399	0.000
Wolff (2011, 2016)	0.006	NA	NA	NA	NA	0.500	0.500	0.500	0.114	0.000

Note. The columns contain: (1) the study name, and in columns (2) to (11) the study-level (unweighted) means of the following variables: (2) the absolute fade-out measure, defined as the difference between a follow-up effect size and its post-test anchor; (3) the effect size based on the raw scores from identical test items to measure progress or decline between post and follow-up tests for the treatment group; (4) the absolute fade-out measure restricted to identical tests; (5) the effect size based on a comparison between a sample outside of the intervention and the treatment group; (6) the effect size based on a comparison between a sample outside of the intervention and the control group; (7) the indicator for the sample being identical at post and follow-up test; (8) the indicator for having a duration between post and follow-up test between 12 and 17 months; (9) the indicator for having a duration between post and follow-up test of more than 17 months; (10) the mean across the post-test effect sizes serving as anchors to the follow-up effect sizes;

and (11) the indicator for the tests being identical at post and follow-up. An entry of NA in the table implies that the mean could not be computed for the study in question (e.g., because of lacking information). Note that indicator variables may vary within studies, and the mean can therefor be between 0 and 1.

Supplemental Material B: Identification of Identical Test Items and Sequential Test of H5a

There were 16 studies included in the primary analysis of H5a and b. Regarding 15 of these, we either received confirmation from a study author that the tests contained identical test items at post and follow-up (Dyson et al., 2013, 2015, 2020; Fälth et al., 2013; Hallstedt et al., 2018; Hatz & Sachse, 2010; Jordan et al., 2012; Konerding et al., 2020; Lindström-Sandahl et al., 2024; Lopez-Pedersen et al., 2023; Powell et al., 2021, 2022; Schüler-Meyer et al., 2019; Svensson et al., 2021; Torgesen et al., 2010), or had access to the original item level data and test material (Hansson, 2015). In cases when it was not obvious from the text, we also asked about the raw scores. In four studies, tests contained identical items but had some features that may complicate the interpretation of the results. We discuss these four studies below and document the information we based the assessment on regarding the one study for which we did not receive author confirmation. After this discussion, we report results from sensitivity analyses that excluded some of the studies with uncertainties.

Dyson et al. (2013, 2015) and Jordan et al. (2012)

We received confirmation from a study author that these three studies used the exact same version of Woodcock–Johnson III Tests of Achievement Form C Brief Battery at post and follow-up: the Applied Problems and Calculations subtest in Dyson et al. (2013) and Jordan et al. (2012), and the Calculations subtest in Dyson et al. (2015). These tests included start- and stop-rules, which may have meant that the students did not answer the identical test items at post and follow-up. However, in all three studies, the follow-up tests were conducted only 8 weeks after the end-of-intervention post-test, and it was unclear how much of a problem start and stop-rules presented. We therefore kept the studies in our primary analysis and present a sensitivity analysis excluding all studies that used a test with start- and stop-rules below.

Schüler-Meyer et al. (2019)

We received confirmation from a study author that the content of the fraction test was identical but the item order and fractions/numbers were changed between post and follow-up. As the test seemed designed to test the exact same skill while avoiding memorization of items, the risk that the test would weight skills differently (i.e., the problem that differences in vertical scaling can induce) seemed low and we included the study in our primary analysis. We present a sensitivity analysis excluding the study below.

Lopez-Pedersen et al. (2023)

We used the Word problems, the Dot comparison, and the Digit comparison tests in this study. The intervention (we included the full 14-week intervention results only) targeted first-grade children and follow-up tests were conducted 6 months after end-of-intervention post-tests. We received confirmation from a study author that the tests used were identical. However, the word-problem test was based on WISC-IV arithmetic tasks, which have a time limit of 30 seconds per item and a stopping rule of four consecutive errors. Furthermore, the Dot and Digit comparisons were "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" (p. 129).

It was unclear how the stopping rule and timed tasks affected how the tests weighted skills at post and follow-up. We included the study in our primary analysis and provide a sensitivity analysis that excluded it below.

Ryder et al. (2008)

We included the effect sizes based on the Burt Word Reading Test and Neale Analysis of Reading Ability tests. The intervention targeted children with reading difficulties in Year 2 and 3 in New Zealand (corresponding to grade 1 and 2 in systems where kindergarten is the entry grade to primary school) and the follow-up was conducted two years after the end-of-intervention post-tests. Ryder et al. (2008) write "All five measures were used as pre- and posttests, and the context-free word recognition [Burt Word Reading Test] and word recognition accuracy [Neale Analysis of Reading Ability] were used as the two-year follow-up measures" (p. 356). The Burt Word Reading Test context-free word recognition subtest consist of "a list of 110 words of increasing difficulty" (p. 357), which children are asked to read. Testing is discontinued when 10 consecutive words were incorrectly read or not attempted. The Neale Analysis of Reading Ability word recognition accuracy asks children to "read aloud a series of short passages that were graded in difficulty" (p. 358). The study mentioned no changes across test occasions to either of these two tests. However, as the follow-up period was long, which in turn may have made the discontinuation rule change the properties of the context-free word recognition subtest, we present a sensitivity test excluding this study below.

Sensitivity Analyses of Identical Test Items

Excluding the studies that used a test with start- and stop-rules (Dyson et al., 2013, 2015; Jordan et al., 2012; Lopez-Pedersen et al., 2023; Ryder et al., 2008) yielded an H5a estimate of 0.39 (95% CI = [0.10, 0.68]) and an H5b estimate of -0.18 (95% CI = -0.36, 0.01]). The H5a

estimate is thus very close to the primary analysis-estimate. Although the latter estimate is not significant on the 95%-level, it is very close and the estimate itself is virtually identical to the primary analysis estimate. Thus, as removing five studies induced a substantial reduction of the sample size, the insignificance is likely a power issue and not an indication that there is no fade-out of effect sizes based on identical test items.

Excluding Schüler-Meyer et al. (2019) yielded the following estimates. H5a: 0.42 (95% CI = [0.20, 0.65]). H5b: -0.15 (95% CI = [-0.29, -0.01]).

We lacked confirmation from authors only regarding Ryder et al. (2008), which is also the study with the longest duration between post and follow-up (24 months). Neither H5a nor H5b was however sensitive to excluding just this study (it was also excluded in the analysis excluding all tests with some form of start- and stop-rules). For H5a, the estimate without Ryder et al. (2008) was 0.36 (95% CI = [0.15, 0.57]) and for H5b, -0.17 (95% CI = [-0.31, -0.03]).

Sequential Test of H5a

The primary analysis test of H5a examined whether treated students forget more than they learn in the period between the post-test and follow-up. As the duration between post and follow-up was relatively long in some studies, it was possible that students had time to both forget and re-learn intervention content. In Figure S1, we present results from a version of the test where we added studies to the meta-regression sequentially by their follow-up duration, starting with the shortest follow-up duration. That is, if two studies had the same follow-up duration they were added to the same meta-regression, which is the reason why there are 10 and not 16 estimates. In Panel a), the x-axis displays the follow-up duration of the latest added studies. In Panel b), the x-axis displays the weighted average duration estimated by the same type of meta-regression we used for the weighted average effect size. In both panels, the y-axis

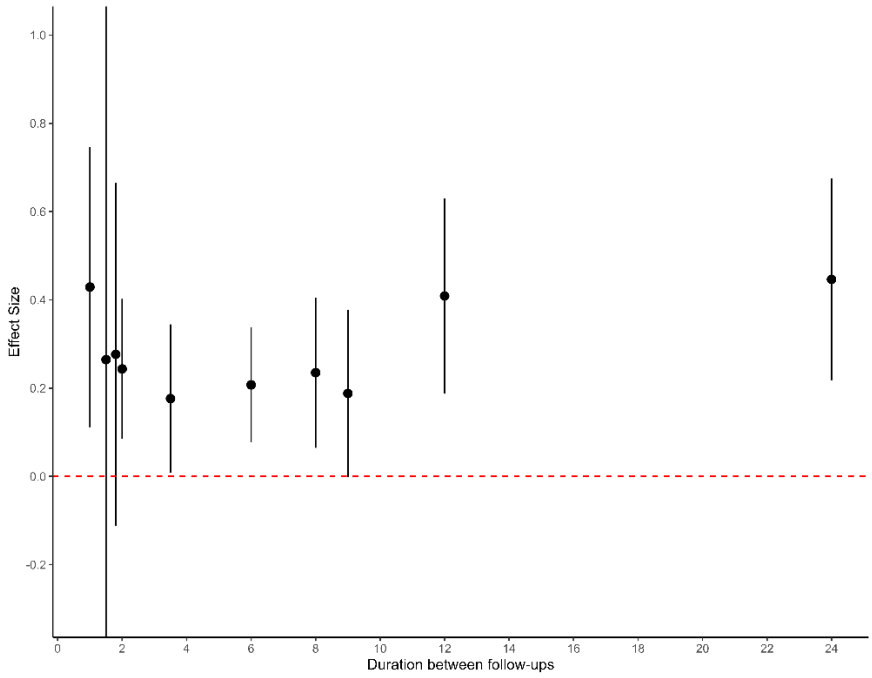
displays the weighted average effect sizes together with their confidence intervals (these estimates are the same in both panels).

The weighted average effect sizes are always positive and relatively large. With two exceptions, they are significant. That is, in our data, there is no duration during which forgetting dominates learning.

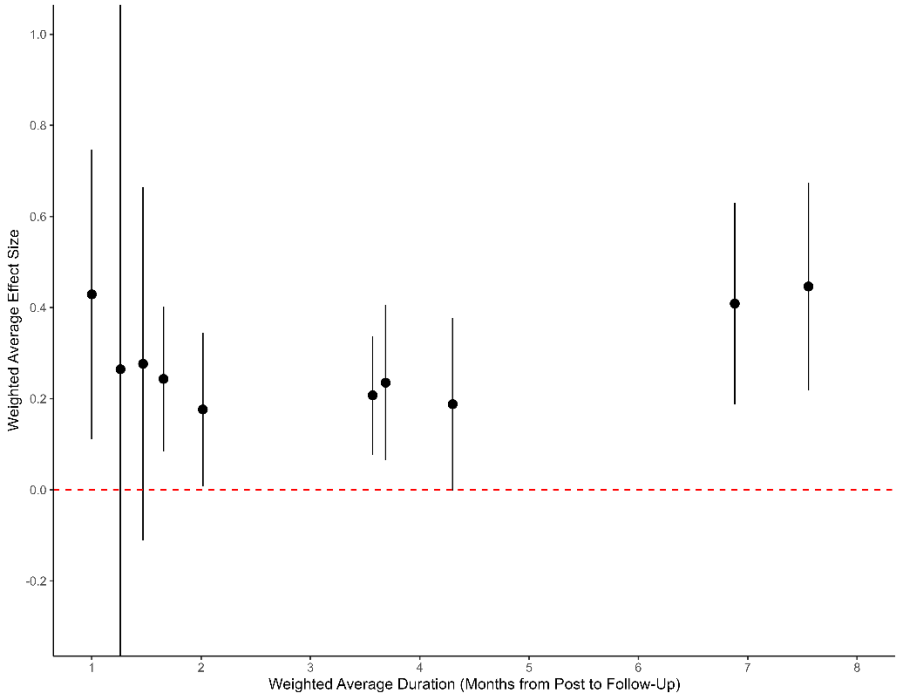
Figure S1

Sequential version of the test of H5a

Panel a)



Panel b)


1.0
0.8
0.6
0.4
0.2
0.0
-0.2
Weighted Average Effect Size
1
2
3
4
5
6
7
8
Weighted Average Duration (Months from Post to Follow-Up)

Supplemental Material C: Blinding of Testers

In the section Measurement Issues, we mentioned that differential blinding of testers may cause fade-out. It is possible that lack of blinding of treatment status may, if testers want the intervention to succeed, inflate post-test effect sizes. If follow-up tests are conducted by testers who are blind to treatment status, this difference may cause fade-out. However, our sample included only one study in which we could ascertain that the blinding of testers differed between the post and follow-up tests (Ryder et al., 2008). Furthermore, running a CHE-RVE model using only effect sizes from studies for which we could ascertain that the blinding was the same for post and follow-up tests yielded a weighted average fade-out of -0.16 (95% CI = [-0.25, -0.07], based on 26 studies and 250 effect sizes). That is, almost identical to the estimate in our primary analysis sample. Differential blinding of testers was therefore not a likely explanation of fade-out in our data.

Supplemental Material D: Full Results of the Sensitivity Analyses

Sample Restrictions Based on Post-Test Effect Sizes

In Table S2, we report results using the specification in column 5 of Table 5 estimated on either the full sample (column 1) or samples which we progressively restricted to treatments with larger post-test effect sizes. We included the primary analysis sample (column marked ES > 0 in the table) for easy reference. For estimates pertaining to H1, H3, and H4, the coefficients were reasonably similar and the outcomes of the tests did not change. For H2, the coefficient on the indicator for post to follow-up durations longer than 17 months became progressively more negative, the more we restricted the sample and was significant for ES > 0.2 and > 0.3.

For H1, we also estimated the weighted average effect size in identical samples. The weighted average effect size was negative and statistically significant in the primary sample ($\hat{\beta}$ = -0.25, 95% CI = [-0.44, -0.07], based on 54 effect sizes and 12 studies). The weighted average effect size was -0.19 in the full sample (95% CI = [-0.39, 0.02], based on 66 effect sizes and 13 studies). With a restriction to ES > 0.1, we obtained the same sample as in the primary analysis. With ES > 0.2 and ES > 0.3, we obtained the same sample, and the weighted average effect size was -0.32 (95% CI = [-0.51, -0.13], based on 40 effect sizes and 10 studies). Thus, larger post-test effect sizes yielded more and more significant fade-out, which may be due to a lower bound at zero. We therefore believe that that the rejection of H1 was robust.

Table S2

Tests of hypothesis 1-4 with different sample restrictions

	Full		ES > 0		ES > 0.1		ES > 0.2		ES > 0.3	
Moderator	Coef.	df	Coef.	df	Coef.	df	Coef.	df	Coef.	df
Intercept	-0.13 [-0.23, -0.03]	12.2	-0.13 [-0.23, -0.03]	11.6	-0.12 [-0.22, -0.01]	12.2	-0.08 [-0.22, 0.05]	10.3	-0.05 [-0.24, 0.14]	8.4
Identical Sample	0.08 [-0.10, 0.26]	9.8	0.06 [-0.12, 0.24]	8.5	0.04 [-0.14, 0.23]	8.2	0.06 [-0.19, 0.30]	8.8	0.03 [-0.25, 0.31]	8.2
Duration (12-17 months)	-0.07 [-0.17, 0.03]	4.6	-0.08 [-0.18, 0.02]	4.2	-0.08 [-0.18, 0.02]	3.9	-0.07 [-0.22, 0.07]	2.8	-0.07 [-0.24, 0.11]	2.3
Duration (>17 months)	-0.12 [-0.26, 0.02]	6.7	-0.13 [-0.27, 0.01]	6.3	-0.14 [-0.29, 0.00]	5.7	-0.19 [-0.35, -0.02]	4.2	-0.21 [-0.40, -0.01]	3.1
Post-test ES	-0.64 [-0.79, -0.49]	8.2	-0.67 [-0.82, -0.52]	7.2	-0.67 [-0.83, -0.52]	6.9	-0.68 [-0.85, -0.51]	5.8	-0.68 [-0.87, -0.47]	5.1
Identical Test	0.00 [-0.10, 0.11]	6.8	0.02 [-0.11, 0.15]	5.3	0.03 [-0.12, 0.18]	4.5	0.01 [-0.13, 0.15]	3.7	0.01 [-0.17, 0.19]	3.1
Effect Sizes	374		341		307		249		203	
Studies	42		40		36		30		22	
Between-study SD	0		0		0		0		0.05	
Within-study SD	0.09		0.08		0.09		0.10		0.11	
Total SD	0.09		0.08		0.09		0.10		0.12	

Note. The outcome variable in all specifications is the absolute fade-out measure. The sample is the full sample (column 1) or consists of the treatments with positive average post-test effect sizes larger than 0 (column 2), 0.1 (column 3), 0.2 (column 4), or 0.3 (column 5). The df column contains the small sample adjusted Satterthwaite degrees of freedom. The terms in brackets below the coefficient estimates denotes the 95% confidence interval.

The primary analysis of H5a yielded a weighted average effect size of 0.40 (95% CI = [0.18, 0.61], based on 111 effect sizes and 16 studies). Changing the sample restrictions resulted in the following estimates. The weighted average effect size was 0.45 in the full sample (95% CI = [0.22, 0.67], based on 130 effect sizes and 17 studies). With a restriction to ES > 0.1, the weighted average effect size was 0.41 (95% CI = [0.16, 0.67], based on 99 effect sizes and 14 studies). With ES > 0.2, the weighted average effect size was 0.37 (95% CI = [0.12, 0.63], based on 88 effect sizes and 12 studies). With ES > 0.3, the weighted average effect size was 0.38 (95% CI = [0.10, 0.66], based on 77 effect sizes and 11 studies). Thus, our rejection of H5a was robust to the definition of the primary analysis sample.

The primary analysis of H5b yielded a weighted average effect size of -0.18 (95% CI = [-0.33, -0.04], based on 111 effect sizes and 16 studies). Changing the sample restrictions resulted in the following estimates. The weighted average effect size was -0.17 in the full sample (95% CI = [-0.31, -0.04], based on 130 effect sizes and 17 studies). Restricting the sample to ES > 0.1, the weighted average effect size was -0.20 (95% CI = [-0.37, -0.04], based on 99 effect sizes and 14 studies). With ES > 0.2, the weighted average effect size was -0.22 (95% CI = [-0.39, -0.05], based on 88 effect sizes and 12 studies). With ES > 0.3, the weighted average effect size was -0.22 (95% CI = [-0.41, -0.04], based on 77 effect sizes and 11 studies). Thus, the rejection of H5b was robust to the definition of the primary analysis sample.

Changing the sample restrictions did not affect the estimates of the follow-up indicator either in the test of H6a or H6b. For H6a, the coefficient on the follow-up indicator was -0.12 in the primary analysis (95% CI = [-0.25, 0.01], based on 325 effect sizes and 18 studies) and -0.12 in the full sample (95% CI = [-0.24, 0.01], based on 354 effect sizes and 19 studies), -0.13 with ES > 0.1 (95% CI = [-0.27, 0.00], based on 301 effect sizes and 18 studies), -0.13 with ES > 0.2

(95% CI = [-0.31, 0.04], based on 236 effect sizes and 16 studies), and -0.12 with ES > 0.3 (95% CI = [-0.37, 0.14], based on 179 effect sizes and 13 studies). The corresponding estimates for H6b were (with the same number of effect sizes and studies as H6a): -0.05 (95% CI = [-0.18, 0.09]), -0.06 (95% CI = [-0.20, 0.08]), -0.05 (95% CI = [-0.20, 0.09]), -0.02 (95% CI = [-0.18, 0.14]), and 0.019 (95% CI = [-0.20, 0.23]). All estimates were thus relatively close to the primary analysis, and none was significant. In turn, the failure to reject H6a and H6b was robust to changing the definition of the sample.

Different Values of ρ

This section contains estimates of the primary analysis using lower and higher values of the correlation between pairs of effect sizes from the same study ($\rho \in \{0,0.2,0.4,0.8\}$). None of the hypothesis tests were sensitive to changing the value of ρ. We rejected or failed to reject exactly the same hypotheses, and the estimates were in all cases close to the ones we obtained in the primary analysis. See Figures S2 to S5 below for the results.

Figure S2

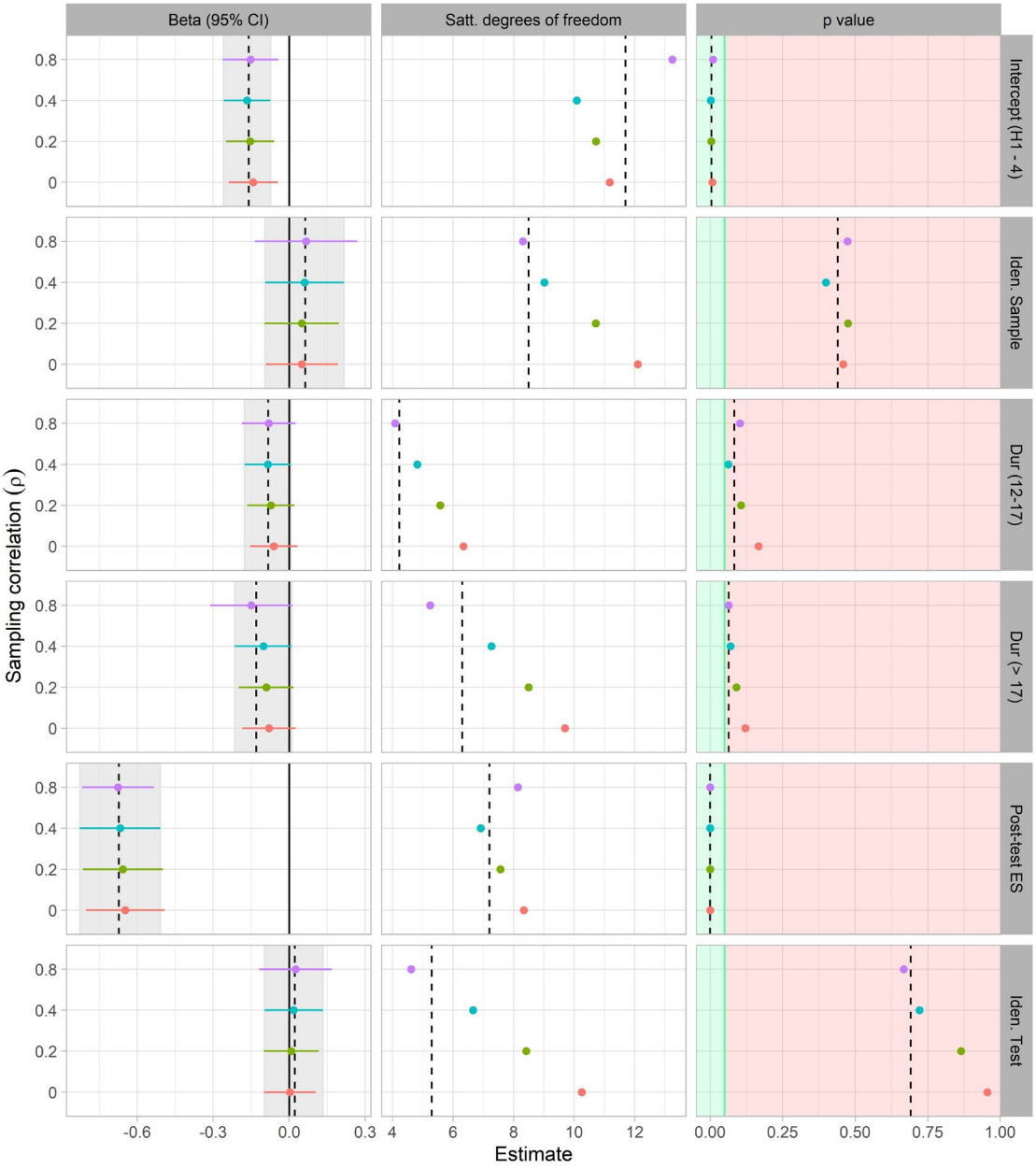
Impact of the assumed sampling correlation on estimates in Table 5 (columns 1 - 4)



Note. Value of ρ on the y-axis. Dashed lines indicate the estimated value when the sampling correlation is 0.6. Gray shades in the first column mark the estimated confidence interval when the sampling correlation is 0.6. The red color in the third column indicates that the null hypothesis is not rejected, and the green color indicates by contrast that it is rejected.

Figure S3

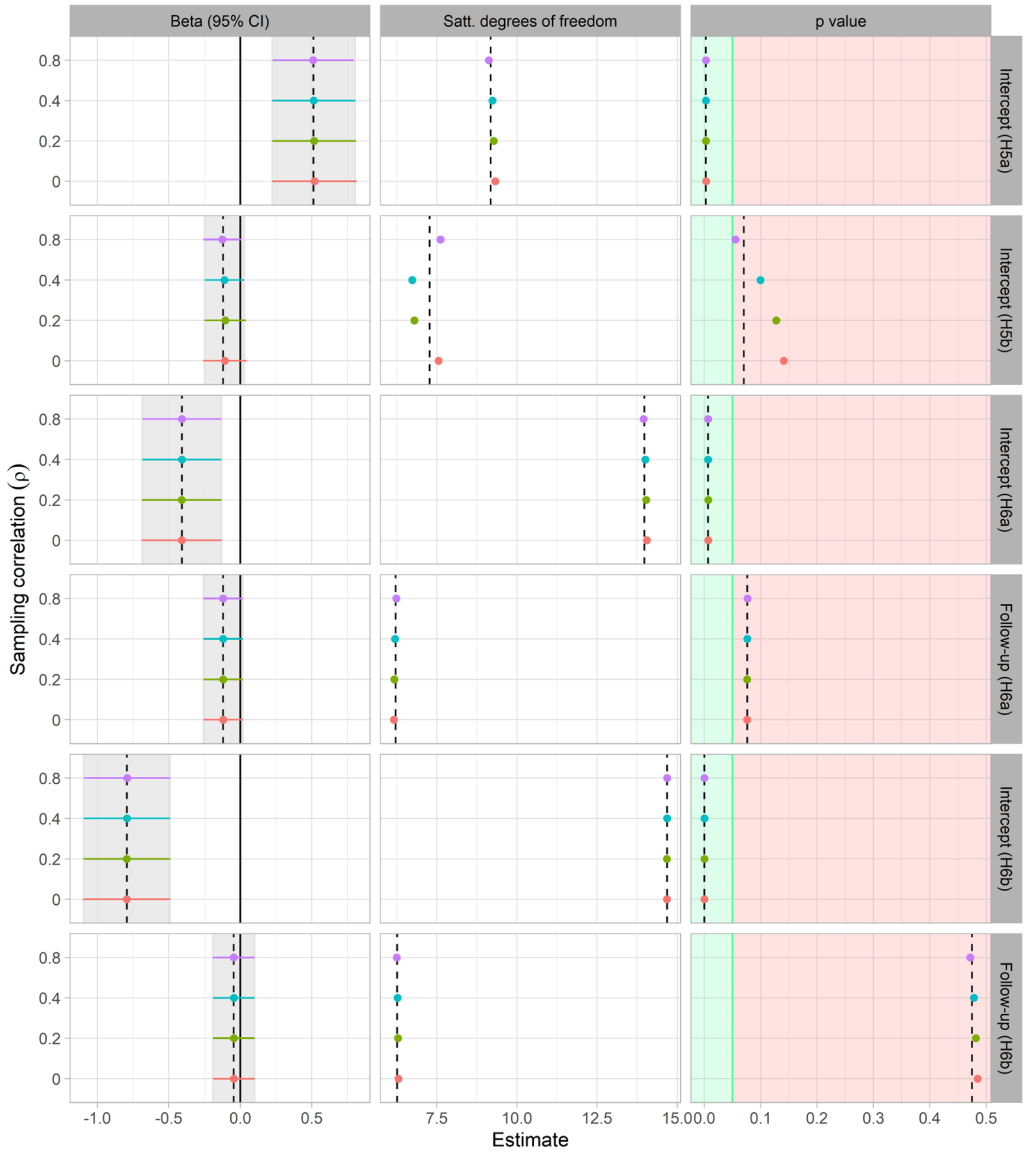
Impact of the assumed sampling correlation on estimates in full model (column 5)



Note. Value of ρ on the y-axis. Dashed lines indicate the estimated value when the sampling correlation is 0.6. Gray shades in the first column mark the estimated confidence interval when the sampling correlation is 0.6. The red color in the third column indicates that the null hypothesis is not rejected, and the green color indicates by contrast that it is rejected.

Figure S4

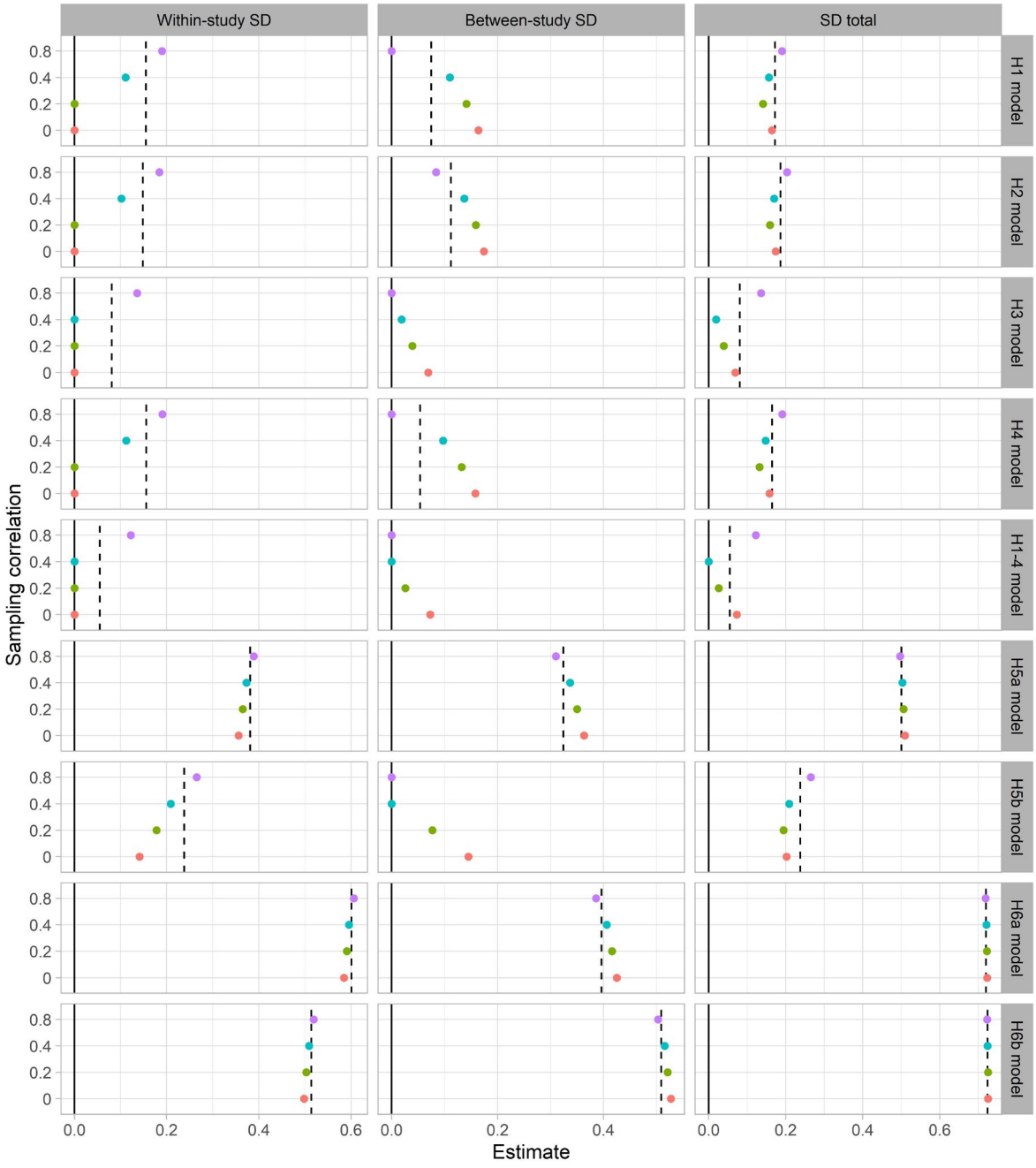
Impact of the assumed sampling correlation on estimates in Table 5 (columns 1 - 4)



Note. Value of ρ on the y-axis. Dashed lines indicate the estimated value when the sampling correlation is 0.6. Gray shades in the first column mark the estimated confidence interval when the sampling correlation is 0.6. The red color in the third column indicates that the null hypothesis is not rejected, and the green color indicates by contrast that it is rejected.

Figure S5

Impact of the assumed sampling correlation on variance estimation across all models.



Note. Value of ρ on the y-axis. Dashed lines indicate the estimated value when the sampling correlation is 0.6.

Identical Samples

Although we found fade-out also in identical samples and insignificant differences between identical and non-identical samples, the differences were not precisely zero. Therefore, changes to the sample composition might to some extent influence our results. We examined this issue by estimating the models for H2-H6 on identical samples.

Table S3 reports the results. As indicated by the low degrees of freedom, many specifications contained too few studies to estimate the coefficients with enough precision. It is thus noteworthy that the post-test effect size is still of similar size and significant in column 1. We could only include one duration coefficient as there were no observations with a longer than 17-month post to follow-up duration. This coefficient is positive instead of negative as in the primary analysis but still insignificant and with a wide confidence interval. Similarly, the coefficient on the identical test-indicator is positive instead of negative but with a very wide confidence interval. The intercept in column 2 (testing H5a) is smaller than in the primary analysis and not significant but still positive. The intercept in column 3 (testing H5b) is more negative than in the primary analysis but not significant. The coefficient on the follow-up indicator in column 4 (testing H6a) is similar in both sign and magnitude to the primary analysis (and not significant), whereas the follow-up indicator in column 5 (testing H6b) is more negative and significant.

Table S3

Tests of hypothesis 2 - 6 using only identical samples

	(1) H2 - 4		(2) H5a		(3) H5b		(4) H6a		(5) H6b	
Moderator	Coef.	df	Coef.	Df	Coef.	df	Coef.	df	Coef.	df
Intercept	-0.07 [-0.59, 0.45]	1.8	0.13 [-0.19, 0.45]	5.3	-0.24 [-0.59, 0.11]	5.3	-0.73 [-1.41, -0.06]	4.7	-0.81 [-1.73, 0.12]	4.8
Duration (12-17 months)	0.06 [-0.99, 1.1]	1.5								
Post-test ES	-0.63 [-1.25, -0.01]	2.4								
Identical test	-0.08 [-0.52, 0.37]	3								
Follow-up							-0.09 [-0.26, 0.09]	1.5	-0.33 [-0.50, -0.17]	1.5
Effect sizes	54		33		33		40		40	
Studies	12		7		7		6		6	
Between-study SD	0.07		0.17		0.19		0.50		0.76	
Within-study SD	0.17		0.36		0.41		0.40		0.54	
Total SD	0.18		0.4		0.45		0.64		0.93	

Note. The outcome variable in column H2-H4 is the absolute fade-out measure. In column H5a, the outcome variable is the ES^I measure. In column H5b, the outcome variable is the fade-out measure, calculated using only raw scores from identical test items. In columns H6a and H6b, the outcome variable is ES^{norm}, calculated for the treatment and control group respectively. The samples in all columns restricts the primary analysis sample to effect sizes based on identical samples. The df column contains the small sample adjusted Satterthwaite degrees of freedom. The terms in brackets below the coefficient estimates denotes the 95% confidence interval.

CE-RVE Model

This section reports the results of running the analysis of the primary samples using the CE-RVE model instead of the CHE-RVE model. We estimated the CE-RVE model using the *robumeta* package for R (Fisher et al., 2017). Table S4 reports the results for H1-H4. Although some of the estimates in Table S4 are different compared to Table 5 in the main text, none of the outcomes of the hypothesis tests have different implications for the mechanisms.[2] Thus, the tests of H1-H4 were robust to using the CE-RVE model instead of CHE-RVE.

Table S5 reproduces the tests of H5 and H6. That is, it corresponds to Table 6 in the main text. One hypothesis test turned out differently with the CE-RVE model. As the coefficient on the follow-up indicator was negative and significant we rejected H6a, which we could not reject in the primary analysis. However, the estimate was close to the one we obtained in the primary analysis (-0.13 vs. -0.12). Thus, the estimates obtained with the CE-RVE model were qualitatively similar to the ones from the CHE-RVE model.

[2] The coefficient on the indicator for identical samples is negative and significant in column 1. However, this result would imply more fade-out in identical samples, i.e., stronger evidence against the sample composition mechanism. The result is not robust though, including other moderators results in a small and insignificant coefficient in column 5.

Table S4

Tests of hypothesis 1-4 using the CE-RVE model

	(1) H1		(2) H2		(3) H3		(4) H4		(5) H1-4	
Moderator	Coef.	df	Coef.	df	Coef.	df	Coef.	df	Coef.	df
Intercept	-0.12 [-0.18, -0.07]	22.4	-0.21 [-0.32, -0.09]	16.5	-0.18 [-0.22, -0.14]	16.6	-0.14 [-0.21, -0.07]	19.5	-0.15 [-0.24, -0.06]	5.6
Identical Sample	-0.15 [-0.30, -0.01]	15.7							-0.01 [-0.14, 0.11]	11.0
Duration (12-17 months)			0.08 [-0.04, 0.2]	24.3					-0.06 [-0.16, 0.03]	10.2
Duration (>17 months)			0.10 [-0.05, 0.25]	15.2					-0.03 [-0.14, 0.09]	8.8
Post-test ES					-0.52 [-0.73, -0.31]	8.3			-0.55 [-0.75, -0.34]	8.8
Identical Test							-0.06 [-0.2, 0.07]	26.2	0.02 [-0.1, 0.13]	11.8
Effect Sizes	341		341		341		341		341	
Studies	40		40		40		40		40	
Between-study SD	0.15		0.17		0		0.16		0	

Note. The outcome variable in all specifications is the absolute fade-out measure. The sample consists of the treatments with positive average post-test effect sizes. All specifications use the CE-RVE model. The df column contains the small sample adjusted Satterthwaite degrees of freedom. The terms in brackets below the coefficient estimates denotes the 95% confidence interval.

Table S5

Tests of hypotheses 5a, 5b, 6a, and 6b using the CE-RVE model

	(1) H5a		(2) H5b		(3) H6a		(4) H6b	
	Coef	df	Coef	Df	Coef	df	Coef	df
Intercept	0.40 [0.14, 0.66]	13.9	-0.18 [-0.29, -0.07]	13.4	-0.37 [-0.57, -0.17]	15.7	-0.75 [-1, -0.5]	15.7
Follow-up					-0.13 [-0.23, -0.02]	16.9	-0.02 [-0.14, 0.09]	16.9
Effect Sizes	111		111		325		325	
Studies	16		16		18		18	
Between-Study SD	0.51		0.32		0.50		0.60	

Note. The specifications in column 1 (H5a) and 2 (H5b) used the sample of effect sizes based on identical test items. The specifications in column 3 (H6a) and 4 (H6b) used the sample of effect sizes in which the treatment (column 3) and control group (column 4) were compared to a sample of students outside of the intervention as a reference group. For hypothesis 5a, the outcome variable is the ES^I measure. For hypothesis 5b, the outcome variable is F^{abs} calculated using only raw scores from identical test items. For H6a and H6b, the outcome variable is ES^{norm} calculated for the treatment and control group respectively. All specifications use the CE-RVE model. The df column contains the small sample adjusted Satterthwaite degrees of freedom. The terms in brackets below the coefficient estimates denotes the 95% confidence interval.

Additional Moderators

In this section, we tested whether the results in the primary analysis were robust to adding three additional moderators: the grade the intervention was implemented in, the duration of the intervention, and whether the implementers were school personnel or staff hired by the research study. As shown by the below discussion, the variables are ambiguously related to fade-out and to the mechanisms, which is why they were not included in our primary analysis.

The grade of the intervention may affect fade-out by, for example, restricting the follow-up period or by interventions in earlier grades having a systematically different content (e.g., more, or less focus on constrained skills). The connection to fade-out mechanisms was therefore ambiguous, and we have no clear prediction of the direction of the association between fade-out and grade. To examine the association, we included a (mean-centered) variable in our meta-regressions measuring the average grade the intervention was implemented in.

As students get more training, longer interventions may result in more durable changes to skills. However, longer interventions may also produce intervention fatigue and stigma, which may lead to less responsiveness to further instruction and interventions, and more fade-out.[3] We examined the association with fade-out using a (mean-centered) variable measuring the intervention duration in weeks.

If school personnel conduct the intervention, they may form longer-lasting emotional bonds to the students, which may imply less fade-out through, for instance, improved student motivation. However, research staff may adhere more closely to the intervention content and thereby produce longer-lasting effects. We included an indicator equal to one when school staff implemented the intervention.

We first added the moderators one-by-one to the specification reported in column 5 of Table 5 in the main text, and then added all three in the same meta-regression. The results are shown in Table S6. None of the moderators were significant and the changes to the other estimates were very small. Therefore, outcomes of the hypothesis tests were robust to adding these three moderators.

[3] Many of the included interventions were pull-out tutoring interventions, which have been hypothesized to increase stigmatization of at-risk students (Marston, 1996). Empirically, the duration of the intervention is not always associated with larger post-test effect sizes (see e.g., Dietrichson et al., 2017, 2021).

Estimations Using Only Within-Study Variation

A way to control for study-level confounders is to only use within-study variation in the estimation. We implemented such a specification by including a dummy for each study and dropping the study-level random-effect term. As mentioned in the main text, there was not enough variation in the data to estimate all variables of interest. We therefore only used this specification to test H2, H3, and H4. We used the specification in column 5 of Table 5 in the main text as the starting point but removed the Identical samples-indicator, due to insufficient variation.

The coefficients on the duration-indicators were reasonably close to the ones we obtained in the primary analysis both in terms of magnitude and significance ($\hat{\beta}_{12-17}$= -0.12, 95% CI = [-0.29, 0.05], $\hat{\beta}_{>17}$= -0.15, 95% CI = [-0.35, 0.04]), and so was the coefficient on the post-test effect size ($\hat{\beta}_{ES}$= -0.69, 95% CI = [-0.85, -0.53]). The coefficient on identical tests was more positive (i.e., indicating less fade-out on identical tests) but the confidence interval is very wide ($\hat{\beta}_{test}$= 0.12, 95% CI = [-1.0, 1.2]). Although it is noteworthy that the estimates were so similar to the primary analysis and the outcomes of the hypothesis tests the same, the results should be viewed with some caution due to the effective variation used in the estimations being low. An indication of this problem is that the adjusted degrees of freedom were small for all three variables (1.7, 2.6, 5.4, and 1.4, respectively).

Table S6

Tests of hypothesis 1 - 4 with added moderators

Moderator	(1) Grade		(2) Duration		(3) School staff		(4) All	
	Coef	df	Coef	df	Coef	df	Coef	df
Intercept	-0.12 [-0.22, -0.03]	11.6	-0.13 [-0.23, -0.03]	13.0	-0.11 [-0.27, 0.05]	13.2	-0.12 [-0.27, 0.04]	13.4
Identical sample	0.07 [-0.11, 0.25]	8.4	0.07 [-0.11, 0.25]	8.5	0.06 [-0.16, 0.27]	7.7	0.06 [-0.14, 0.27]	7.5
Duration (12-17 months)	-0.08 [-0.18, 0.02]	4.2	-0.09 [-0.18, 0.01]	4.0	-0.08 [-0.18, 0.02]	3.8	-0.08 [-0.18, 0.03]	3.5
Duration (>17 months)	-0.13 [-0.27, 0.01]	6.3	-0.14 [-0.29, 0.01]	5.6	-0.13 [-0.27, 0.01]	6.0	-0.13 [-0.29, 0.02]	5.4
Post-test ES	-0.68 [-0.83, -0.52]	7.2	-0.67 [-0.82, -0.52]	7.2	-0.67 [-0.82, -0.52]	7.2	-0.67 [-0.83, -0.52]	7.2
Identical test	0.02 [-0.11, 0.14]	5.3	0.03 [-0.11, 0.16]	5.2	0.02 [-0.11, 0.15]	5.3	0.02 [-0.11, 0.15]	5.2
Grade	-0.01 [-0.03, 0.003]	3.1					-0.01 [-0.02, 0.003]	3.3
Duration of intervention			0.0003 [-0.002, 0.003]	2.3			0.0002 [-0.002, 0.002]	3.0
School staff					-0.02 [-0.14, 0.11]	15.4	-0.02 [-0.14, 0.11]	13.2
Effect sizes	341		341		341		341	
Studies	40		40		40		40	
Between-Study SD	0		0		0		0	
Within-Study SD	0.08		0.08		0.08		0.08	
Total SD	0.08		0.08		0.08		0.08	

Note. The outcome variable in all specifications is the absolute fade-out measure. The sample consists of the treatments with positive average post-test effect sizes. The df column contains the small sample adjusted Satterthwaite degrees of freedom. All specifications use the CHE-RVE model. The terms in brackets below the coefficient estimates denotes the 95% confidence interval.

Relative Fade-Out Measure

The implications of the fade-out mechanisms in the main text were stated with an absolute fade-out measure in mind. Here, we first discuss by hypothesis whether the implications are the same for a relative fade-out measure and whether they depend on functional form, and then proceed to the hypothesis tests using the relative measure. Before presenting the results, we discuss some of the issues with using a relative measure more in depth than in the main text. We focus on hypotheses H1-H4 and H5b, for which we used the absolute fade-out measure as the outcome variable in the primary analysis.

Summing up the implications by hypothesis, only the implications of H3 change, because all other implications or predictions are either directional, not about more or less fade-out, or in the case of H2, make a prediction over time.

H1 (Sample Composition). *There is no or less fade-out in studies where the sample is identical at the post and follow-up tests.* This hypothesis concerns only the sample composition mechanism. The implication of this mechanism does not change for a relative measure. That is, if we find fade-out with identical samples at post and follow-up tests, then changes to the sample composition are not the only reason for fade-out. If we find less fade-out in these studies than in other studies, then changes to the sample composition may explain some fade-out.

H2 (Follow-Up Duration). *Fade-out is unrelated to the duration between the post and follow-up tests.* The three skill-based mechanisms imply that effect sizes should fade out more, the longer the follow-up period. The sample composition mechanism is compatible with any functional form of the association between effect sizes and the duration between post and follow-up tests. If the RTM mechanism fully explains fade-out, we should expect to see effect sizes revert to their mean at first follow-up and then stay there. If test variances increase over time, we

should expect fade-out to increase over time. It is unclear whether problems with vertical scaling of tests become more severe over time, and this mechanism is compatible with both rejecting and not rejecting H2. None of these implications change with a relative fade-out measure.

H3 (Post-Test Effect Size). *Effect size fade-out is unrelated to the post-test effect size.* The skill decay, compensatory catch-up, treatment-induced (t) and (c), and RTM mechanisms all predicted that fade-out would be increasing in the post-effect size, i.e., a rejection of H3. These implications were stated with an absolute fade-out measure in mind. Below, we show that for simple models of skill development, the implications for the skill-based mechanisms are not necessarily the same for a relative fade-out measure whereas they are the same for RTM. Note that we do not claim that the models used below are the most likely versions of skill development. Our aim is not to develop a fully-fledged model of skill development, just to show in a simple way that the mechanisms' implications for post-test effect sizes may differ between absolute and relative measures of fade-out, and that the implications are sometimes dependent on functional form with a relative measure.

Denote the average skill measure θ_m^G in period $m \in \{p, fu\}$, where p is the post-test and fu is the follow-up, and group $G \in \{T, C\}$, where *T* and *C* denote the treatment and control groups, respectively. Let $\theta_m^G, I_m^G, \delta_\theta^G, \delta_I^G \in [0, \infty)$, where I_m^G is the investment and $\delta_\theta^G, \delta_I^G$ are parameters governing the skill development between the post and follow-up tests in each group and $\delta_\theta^C - \delta_\theta^T \equiv \epsilon_\theta, \delta_I^C - \delta_I^T \equiv \epsilon_I$.

To focus on the skill-based mechanisms and rule out some measurement issues, we make two assumptions:

Assumption (1): the standard deviation used to calculate the effect size is equal across periods, i.e., $\sigma_{fu} = \sigma_p = \sigma \in (0, \infty)$.

Assumption (2): $\theta_p^T > \theta_p^C > 0$.

Assumption (1) rules out the increasing variance mechanism and we invoke assumption (2) to ensure a positive post-test effect size.

Let $\theta_{fu}^G = \delta_\theta^G \theta_p^G + \delta_I^G I_{fu}^G$, where $\delta_I^G > 0$ and I_{fu}^G may, depending on the mechanism, be a function of the post-test skill level.

Define $g_p = \left(\theta_p^T - \theta_p^C\right)/\sigma$ and g_{fu} analogously. Let $F_{abs} = g_{fu} - g_p$ be the absolute fade-out measure and $F_{rel} = g_{fu}/g_p$ the relative fade-out measure. Rewriting these expressions in the following way will be useful:

$$F_{abs} = g_{fu} - g_p = g_p\left(\delta_\theta^T - \delta_I^T - 1\right) - \frac{\epsilon_\theta \theta_p^C}{\sigma} - \epsilon_I I_{fu}^C$$

$$F_{rel} = \frac{g_{fu}}{g_p} = \delta_\theta^T - \delta_I^T - \frac{1}{g_p}\left(\frac{\epsilon_\theta \theta_p^C}{\sigma} + \epsilon_I I_{fu}^C\right)$$

Skill decay assumes that the treatment group skills decreases between post and follow-up, i.e., $\theta_{fu}^T < \theta_p^T$, which causes fade-out. . Since $\delta_I^T I_{fu}^T \geq 0$ by definition, this means that $\delta_\theta^T < 1$ to get fade-out. The mechanism does not work through investments, i.e., we can assume that $\epsilon_I = 0, I_{fu}^T = I_{fu}^C$. Thus, if skill decay is causing fade-out, then F_{abs} is negative and decreasing in g_p (because $\delta_\theta^T < \delta_I^T + 1$).

With these assumptions, then $\frac{\partial F_{rel}}{\partial g_p} = \frac{\frac{\epsilon_\theta \theta_p^C}{\sigma}}{g_p^2} \geq 0$. That is, in this model, the skill decay mechanism can imply either no relationship between the post-test effect size and fade-out (if

$\epsilon_\theta = 0$) or that there will be less fade-out, the larger the effect size (if $\epsilon_\theta > 0$). The prediction regarding H3 is thus ambiguous and depends on functional form assumptions.

When discussing the relation between post-test effect sizes and fade-out in the main text, we mentioned that a potential reason for compensatory catch-up may be that only a proportion of skills taught in the intervention transfer to the skills measured at follow-up (Bailey et al., 2018). In our model, we can represent this as $0 < \delta_\theta^T < 1$, assuming a non-identical test so that a proportional transfer does not have to imply skill decay. Then the implications of the compensatory catch-up mechanism are the same as for skill decay. Because I_{fu}^C and the ϵ_i are zero or positive, the absolute fadeout measure will be decreasing in g_p whereas $\frac{\partial F_{rel}}{\partial g_p} = \frac{\frac{\epsilon_\theta \theta_p^C}{\sigma} + \epsilon_I I_{fu}^C}{g_p^2} \geq 0$. The prediction regarding H3 is thus ambiguous and depend on functional form assumptions.

The treatment-induced catch-up mechanisms explain fade-out by the treatment having negative (treatment-induced t) or positive (treatment-induced c) side-effects on the skill investments in the next period. Thus, in both versions $I_{fu}^T < I_{fu}^C$. We express this idea using a very simple functional form below: $I_{fu}^G = \kappa - \theta_p^G$, where κ is a positive constant. The results are similar with e.g., $I_{fu}^G = \kappa - {\theta_p^G}^2$ or $I_{fu}^G = \kappa - \sqrt{\theta_p^G}$. The treatment-induced catch-up mechanisms does not work through differences in how the earlier skill-level affect the current skill-level or through different returns to investments, so we can set $\epsilon_\theta, \epsilon_I = 0$. Then, $F_{abs} = g_p\left(\delta_\theta^T - \delta_I^T - 1\right)$, which requires $\delta_\theta^T < \delta_I^T + 1$ to get fade-out. In that case, a higher post-test effect size yields more fade-out.

With the same assumptions, F_{rel} simplifies to $F_{rel} = \delta_\theta^T - \delta_I^T$, which is negative if there is fade-out. However, the relative fade-out measure is then unrelated to the post-test effect size. Thus, also for treatment-induced (t) and (c), we can obtain an example in which the implications regarding the post-test effect size for fade-out are different with a relative measure.

To model the implications of RTM in a simple way, assume that the observed effect sizes are determined as: $g_k^m = \bar{g} + \varepsilon_k^m, \varepsilon_k^m \sim N(0, \sigma_\varepsilon^2), 0 < \sigma_\varepsilon^2 < \infty$ for all effect sizes *k* and periods $m \in \{p, fu\}$. This formulation means that the expected value of the post-test effect size is the same for all *k* at both post and follow-up, $\mathbb{E}(g_k^m) = \bar{g}$. To introduce RTM, assume that only effect sizes with anchors such that $g_k^p > k > \min(g^p)$ are observed at follow-up, where k is a cutoff and $\min(g^p)$ is the smallest effect size observed at post-test. This assumption implies that $\mathbb{E}(\hat{g}_k^p) \equiv \hat{g} > \bar{g}$, where $\hat{g}_k^p$ are the anchors to the observed effect sizes at follow-up. Let the absolute and relative fade-out measures be defined as above. Then, $\mathbb{E}(F_{abs}) = \bar{g} - \hat{g} < 0$, which is decreasing in $\hat{g}$. $\mathbb{E}(F_{rel}) = \frac{\bar{g}}{\hat{g}} < 1$, which is also decreasing in $\hat{g}$. Thus, the implication of RTM for H3 is qualitatively the same for both fade-out measures.

H4 (Overlapping Test Domains). *The degree of overlap between the domains tested at post and follow-up is unrelated to fade-out.* This hypothesis is related to theories that predict more transfer between tests of similar or overlapping skills (i.e., that near transfer is more likely than far transfer). Compensatory catch-up would be more likely when there is less overlap in this sense, and this holds also for a relative fade-out measure. No other mechanism has any implication for this mechanism.

H5b (Effect Sizes, Identical Test Items). *If the raw scores from identical test items are used to calculate effect sizes at post and follow-up, there is no fade-out.* This hypothesis concerns

mainly the scaling of tests, which should not be an issue when the raw scores from identical test items are used to calculate effect sizes. If we find evidence of fade-out also on identical test items, such results would be evidence against scaling as the only relevant fade-out mechanism. This would be true also for a relative measure.

Results With a Relative Fade-Out Measure

As mentioned in the main text, there are additional issues with using a relative fade-out measure instead of an absolute measure. These issues have consequences for both the samples used and the functional form of the measure. We discuss these issues and consequences in turn below, before we present the results.

First, a relative measure, such as g_{kfs}/g_{kps}, is a ratio and therefore not defined if the post-test effect size is zero. There were two effect sizes where $g_{kps} = 0$ in the primary analyses of H1-H4, which we excluded from the analyses of the relative measure.

Second, for two reasons, a relative measure risks introducing substantial nonlinearities, which may both complicate the interpretation of the results and the estimations. a) Regarding the interpretation, the values of the relative measure may shift around zero in ways that make the measure difficult to interpret. b) Small absolute fade-out may seem very large with a relative measure if the post-test effect size is small, and consequently, we may obtain a large number of outliers.

Some examples may illustrate a) and b). Suppose first that g_{kfs}= -0.01 and g_{kps}= 0.01, which yields absolute fade-out of F_{ks}= -0.02 and relative fade-out g_{kfs}/g_{kps} = -1. Suppose then that g_{kfs}= 0.01 and g_{kps}= 0.02, which yields F_{ks}= -0.02 but g_{kfs}/g_{kps} = 0.33, which is substantially smaller and of a different sign, despite the absolute fadeout being exactly the same.

If we instead have growing absolute effects, these effects may show up as negative values of the relative measure. Compare g_{kfs}= -0.01 and g_{kps}= -0.03 with g_{kfs}= 0.01 and g_{kps}= -0.01. In both situations, we have growing effects of the same absolute magnitude (F_{ks}= 0.02), but the relative measure is 0.33 (wrongly indicating a smaller effect size at follow-up as the ratio is below one) in the first case and -1 (large and with the wrong sign) in the second case.

Due to the problems of negative values, we believed that excluding effect sizes with negative values from the analyses yielded more interpretable results (below, we provide results with negative effect sizes included as well).

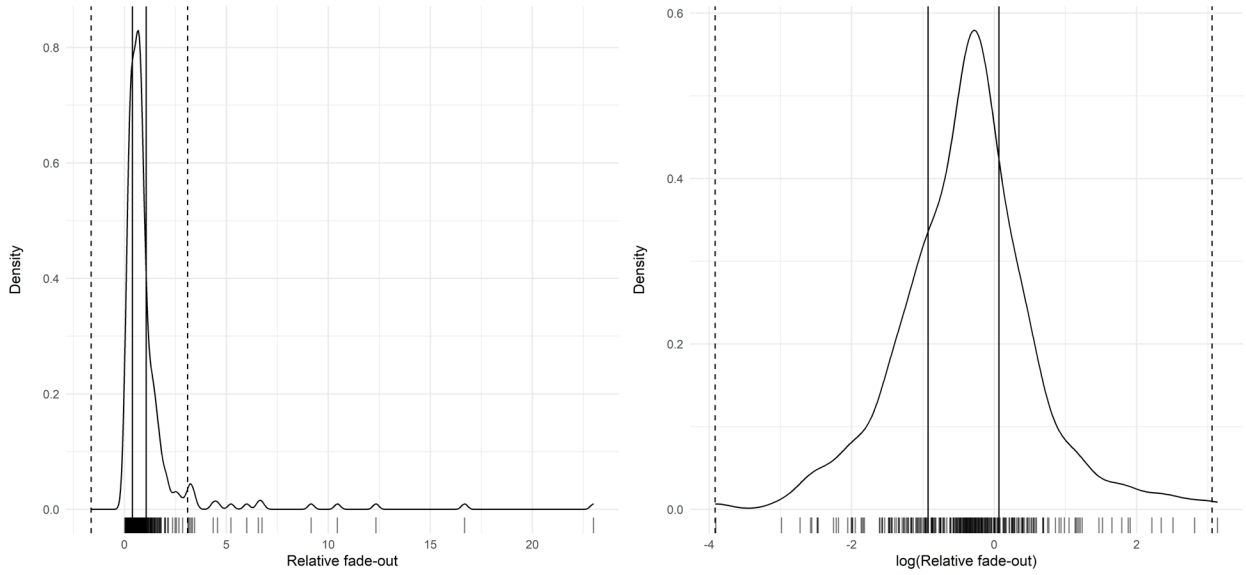
The examples also indicated the outlier-problem, which was clearly present in our data. The left panel of Figure S6 shows the density of g_{kfs}/g_{kps}, estimated on a sample which took the H1-H4 primary analysis sample as the starting point but excluded post-test effect sizes equal to zero and any effect size with negative values. The density has a long right tail with many and large outliers (when we included negative effect sizes, the density has a similarly long left tail), and it is not even approximately normally distributed. To mitigate this problem, we defined a new relative measure $F_{ks}^{rel} = \log(g_{kfs}/g_{kps})$, where log( ) is the natural logarithm. The right panel of Figure S6 shows the density of F_{ks}^{rel}, estimated on the same sample as the density in the left panel. There are clearly fewer outliers and the density is a closer approximation to the normal. We therefore proceeded to estimate H1-H4 and H5b with this measure.

Table S7 presents the results. Column 1 (H1-H4) and 4 (H5b) used the sample excluding all negative effect sizes as well as post-test effect sizes equal to zero and with F_{ks}^{rel} as the outcome measure–our preferred specification with the relative measure. For reference, we also included column 2 (H1-H4) and 5 (H5b), which used the same sample but with g_{kfs}/g_{kps} as the

outcome measure (i.e., outliers and the non-normal distribution may cause more problems), and column 3 (H1-H4) and (H5b), which used the primary analysis sample for the respective hypothesis but with only post-test effect sizes exactly equal to zero excluded and g_{kfs}/g_{kps} as the outcome measure (i.e., negative effect sizes, outliers, and the non-normal distribution may cause problems). Note that we omitted the small sample adjusted Satterthwaite degrees of freedom from the table for layout reasons. There was no coefficient with degrees of freedom lower than 5.6.

Figure S6

Distributions of relative fade-out measures



Note: The left panel shows the density of the relative measure defined as g_{kfs}/g_{kps} and the right panel shows the density of the relative measure defined as $\log(g_{kfs}/g_{kps})$. The solid lines shows the interquartile range and the dashed lines shows three times the interquartile range.

Table S7

Tests of hypotheses H1-H4 and H5b using relative measures of fade-out

	(1) H1-H4	(2) H1-H4	(3) H1-H4	(4) H5b	(5) H5b	(6) H5b
Intercept	-0.001 [-0.42, 0.41]	2.1 [1.2, 3.0]	1.7 [0.66, 2.6]	-0.18 [-0.60, 0.24]	1.8 [0.93, 2.7]	0.96 [0.26, 1.7]
Identical sample	-0.20 [-0.80, 0.39]	-0.75 [-1.6, 0.07]	-1.3 [-2.4, -0.29]			
Duration (12-17 months)	-0.43 [-0.76, -0.09]	-1.1 [-2.1, -0.20]	-1.2 [-2.3, -0.13]			
Duration (>17 months)	-0.70 [-1.1, -0.31]	-1.2 [-2.1, -0.26]	-1.0 [-2.0, -0.13]			
Post-test ES	-1.1 [-1.7, -0.58]	-2.0 [-3.3, -0.78]	-0.24 [-1.3, 0.83]			
Identical test	0.28 [-0.31, 0.87]	0.67 [-0.04, 1.4]	0.30 [-0.17, 0.76]			
Effect Sizes	284	284	339	83	83	109
Studies	38	38	38	15	15	16
Between-Study SD	0.42	0.0001	0.0001	0.42	0.0004	0.0002
Within-study SD	0.78	1.9	2.4	1.1	3.3	3.7
Total SD						

Note. All samples are restricted to treatments with positive average post-test effect sizes and post-test effect sizes not equal to zero. In column 1 and 4, we furthermore removed all negative effect sizes. The specifications in column 4-6 (H5b) used the sample of effect sizes based on identical test items. The outcome variable is F^{rel} in column 1 and 4, and g_{kfs}/g_{kps} in the other columns. The terms in brackets below the coefficient estimates denotes the 95% confidence interval.

For H1-H4, the results in column 1 are very similar to the primary analysis in terms of the signs of the coefficients and reasonably similar in terms of significance. The main difference is that the duration-variables are both significant with the relative measure. The results are similar also in column 2 and 3, but the post-test effect size loses its significance in column 3 and the estimates for the duration show signs of non-monotonicity. Regarding H5b, the coefficient is

negative in column 4 but, in contrast to the primary analysis, not significant. The results in column 5 and 6 are further off from the primary analysis (column 5 even indicates large growing effects). As mentioned, we believe there were strong reasons to prefer the specifications in column 1 and 4.

Samples Restricted to RCTs of Tutoring Interventions

Table S8 and S9 presents the results from analyses with the same specifications as in the primary analyses but using samples restricted to RCTs and tutoring interventions. The results were overall very similar both in terms of the size of coefficients and the significance. There was only one outcome of a hypothesis test that changed: the test of H5b is no longer significant. However, note that the coefficient is actually larger in the RCT/tutoring sample than in the primary sample (-0.19 vs. -0.18). Because the RCT/tutoring sample contains 5 studies less–a substantial reduction compared to the 16 studies in the primary analysis sample–the lack of significance is likely a power issue.

Table S8

Tests of hypotheses 1-4 restricting data to include RCTs and tutoring interventions only

	(1) H1		(2) H2		(3) H3		(4) H4		(5) H1-4	
Moderator	Coef.	df	Coef.	df	Coef.	df	Coef.	df	Coef.	df
Intercept	-0.19 [-0.32, -0.06]	18.2	-0.16 [-0.29, -0.02]	15.7	-0.17 [-0.22, -0.12]	19.1	-0.19 [-0.31, -0.06]	18	-0.15 [-0.29, -0.01]	11.1
Identical Sample	0.03 [-0.31, 0.37]	6.5							0.04 [-0.17, 0.25]	7.1
Duration (12-17 months)			0 [-0.19, 0.19]	5.3					-0.04 [-0.21, 0.12]	4.9
Duration (>17 months)			-0.1 [-0.36,0.17]	4.8					-0.12 [-0.34, 0.11]	4.2
Post-test ES					-0.66 [-0.88, -0.44]	6.2			-0.67 [-0.88, -0.46]	6.1
Identical Test							0.03 [-0.26, 0.32]	5.4	0.03 [-0.16, 0.22]	3.7
Effect Sizes	267		267		267		267		267	
Studies	29		29		29		29		29	
Between-study SD	0.12		0.13		0		0.12		0	

Note. The outcome variable in all specifications is the absolute fade-out measure. The sample consists of the treatments with positive average post-test effect sizes. All specifications use the CHE-RVE model. The df column contains the small sample adjusted Satterthwaite degrees of freedom. The terms in brackets below the coefficient estimates denote the 95% confidence intervals.

Table S9

Tests of hypotheses 5a, 5b, 6a, and 6b restricting data to include RCTs and tutoring interventions only

	(1) H5a		(2) H5b		(3) H6a		(4) H6b	
	Coef	df	Coef	df	Coef	df	Coef	df
Intercept	0.41 [0.12, 0.69]	9.5	-0.19 [-0.39, 0.02]	8.8	-0.38 [-0.67, -0.09]	13.3	-0.79 [-1.11, -0.47]	13.8
Follow-up					-0.13 [-0.27, 0.01]	6.3	-0.05 [-0.19, 0.10]	6.4
Effect Sizes	57		57		313		313	
Studies	11		11		15		15	
Between-Study SD	0.32		0.12		0.41		0.53	

Note. The specifications in column 1 (H5a) and 2 (H5b) used the sample of effect sizes based on identical test items. The specifications in column 3 (H6a) and 4 (H6b) used the sample of effect sizes in which the treatment (column 3) and control group (column 4) were compared to a sample of students outside of the intervention as a reference group. For hypothesis 5a, the outcome variable is the ES^I measure. For hypothesis 5b, the outcome variable is F^{abs} calculated using only raw scores from identical test items. For H6a and H6b, the outcome variable is ES^{norm} calculated for the treatment and control group respectively. All specifications use the CHE-RVE model. The df column contains the small sample adjusted Satterthwaite degrees of freedom. The terms in brackets below the coefficient estimates denotes the 95% confidence interval.

Supplemental Material E: Assumptions and Calculations of the Regression to the Mean Effect

This section first provides results regarding the selection of studies for follow-up and the test of the combined implications of assumption A1 and A2. The main results from the tests of these assumptions were described and presented in the main test. We then present calculations of the size of the RTM effect under different assumptions.

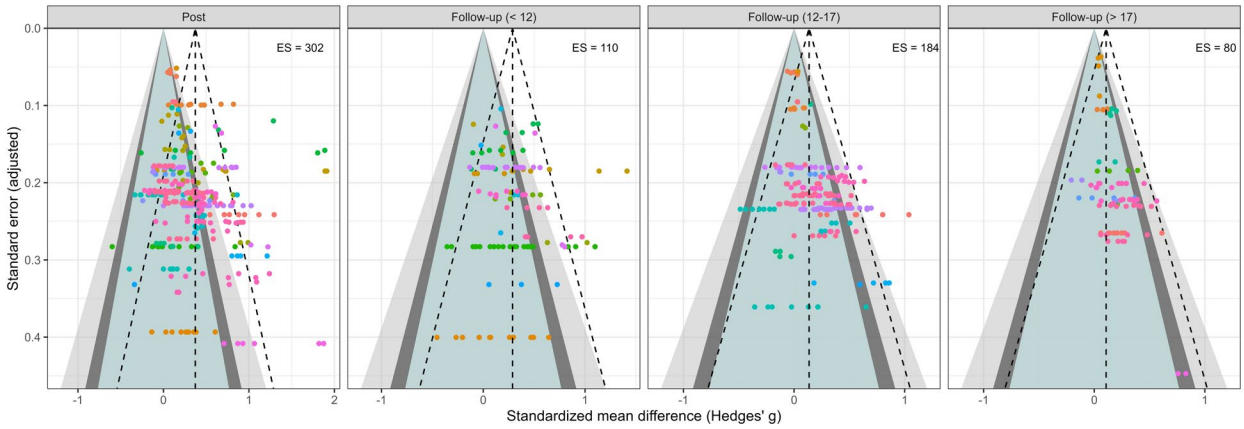
Funnel Plots

Figure S7 shows funnel plots of effect sizes (Hedges' *g* on the x-axis) and the standard errors of the effect sizes (y-axis) at post-test (left-most figure), follow-up less than 12 months, between 12 and 17 months, and more than 17 months after the end of intervention. The dashed center-line shows the weighted average effect size. The green-grey, dark grey, and light grey areas centered around zero shows *p*-values corresponding to 10%, 5%, and 1% significance levels.

There are some signs of asymmetry. In particular at post-test and, especially, at follow-up more than 17 months after end of intervention, the distributions seem to be missing some small studies with relatively small effects. Selection of long-term follow-ups may have biased our estimates of the duration-indicators towards null and may explain the lack of significance in the primary analysis.

Figure S7

Contour-enhanced funnel plots



Note: The plot depicts the relationship between effect size estimates and their corresponding standard errors across the time from the end of the intervention. The colors of the points visualize study affiliations. Following Pustejovsky and Rodgers (2019), we used the adjusted standard errors based on sample sizes (i.e., same as the weights in our meta-regressions) to avoid biasing the plots because of the correlation between the effect size and its regular standard error. Effects in the green region have *p*-values above 0.1, effects in the dark gray area have *p*-values between 0.05-0.1, effects in the light gray area contain *p*-values ranging from 0.01 to 0.05, and finally effects in the white area have *p*-values less than 0.01. We calculated *p*-values using the adjusted standard errors of the effect sizes, which are not necessarily equal to the *p*-values reported by authors of the primary studies. The dashed lines mark the distribution about the mean effect size, estimated via SCE+ model. Effects falling outside the dashed triangle are statistically different from the estimated mean effect size.

Test of Assumption A1 and A2

We examined A1 by estimating the association between having a follow-up and the weighted average effect size and the largest effect size in a subset of the sample of Dietrichson et al. (2021). We excluded waiting list designs from this sample, which are unlikely to have follow-ups because the control group receives the intervention after the post-test. We estimated the association using meta-regressions with an intercept and an indicator for having a follow-up as a single moderator, and all effect sizes or the largest effect size as the outcome variable. The first specification used the CHE-RVE model, and the latter a regular meta-regression estimated with restricted maximum likelihood (each study contributed one effect size in this specification).

We found positive associations between the follow-up indicator and both the weighted average effect size ($\hat{\beta}$ = 0.13, 95% CI = [-0.015, 0.27], based on 897 effect sizes and 162 studies), and the largest effect size ($\hat{\beta}$ = 0.24, 95% CI = [0.041, 0.43], based on 162 effect sizes and the same number of studies). The latter association was significant.

Another potential implication of the RTM assumptions (A1 and A2) is that conditional on the post-test effect size, effect sizes with larger standard errors are more likely to be over-estimated and should fade out more. We tested this implication using the following specification:

$$F_{ks} = \alpha + \beta_1 g_{kps} + \beta_2 SE_{g_{kps}} + \beta_3 \left(g_{kps} \times SE_{g_{kps}} \right) + \mu_s + \upsilon_{ks} + \varepsilon_{ks}$$

where F_{ks} is the absolute fade-out measure of effect size k in study s, g_{kps} is the (mean-centered) post-test effect size of the anchor of k, $SE_{g_{kps}}$ is the (mean-centered) standard error of g_{kps}, and μ_s, υ_{ks}, and ε_{ks} are defined as in the tests of H1-H4. The main parameter of interest is β_3. If negative, effect sizes with both large effects and large standard errors fade out more, in line with the assumptions.

Contrary to the hypothesized implication of the assumptions, β_3 was positively associated with fade-out, albeit not significant (β_3 = 0.43, 95% CI = [-1.3, 2.2]). The standard error was also positively associated with fade-out (β_2 = 0.71, 95% CI = [0.09, 1.3]). These results were not likely to be caused by multicollinearity. Although the post-test effect size and the standard error were correlated, this correlation was around 0.2, which did not seem high enough to cause problems with our sample size (340 effect sizes and 40 studies). Furthermore, it was only conditional on the post-test effect size that the association between the standard error and fade-out was positive. In a regression excluding the post-test effect size, the coefficient on the standard error was -1.03 (95% CI = [-2.0, -0.12]).

Thus, we found no support for the implication of the combination of A1 and A2. However, these results did not provide clear-cut evidence against RTM, although they add to the evidence that RTM was not the only fade-out mechanism. When conditioning on the post-test effect size, the standard error is fully determined by the sample size. Another interpretation of the results was therefore that smaller studies fade out less.

It was beyond the scope of this study to provide a full examination of why effects in smaller studies may be more persistent but we see at least four possible explanations: a) smaller studies may, by chance, be more likely to obtain unbalanced treatment and control groups before the intervention. If studies are selected on the post-test effect size and the imbalance is in favor of the treatment group, then this effect size ought to be *less likely* to fade-out. That is, even if the cause of the initial pre-treatment imbalance is random, it is of a type that is likely to persist, and thus not cause RTM. Smaller studies may also b) be more sensitive to selective reporting of both post-tests and later follow-ups (because larger studies can generate a wider range of statistically

significant effects), c) have different study characteristics (e.g., shorter duration between post and follow-up tests or different types of intervention), d) be better implemented, and e) include more motivated schools (e.g., Dietrichson et al., 2021). Better implementation and more motivated schools may improve the persistence of effects.

Of these, we tested a) in the following way (systematically testing b)-e) would be difficult with our data): we calculated an "effect size" at pre-test and then included this variable as the outcome in a regression with an intercept and the mean-centered sample size as a predictor (we did not weight the observations). We could calculate this effect size for 34 studies and obtained 180 observations. We found a very small and not significant association (p = 0.412). The magnitude of the coefficient implied that an increase of the sample size with 100 students increased the pre-test effect size with 0.01. Reassuringly, the intercept was also small, -0.02; that is, there was no indication of large imbalances between the treatment and control groups at pre-test.

Calculations of the Size of the Regression to the Mean Effect

Below, we examined how much of the fade-out we observed that might be explained by RTM. The calculations relied on several approximations and assumptions, so the results should be interpreted cautiously.

We used the following formula for the RTM effect as our starting point (Barnett et al., 2005; Linden, 2013):

$$RTM = \sigma(1 - \rho_{pf})C(z)$$

where σ is the total effect size standard deviation at post-test, ρ_{pf} is the correlation between the post-test and the follow-up effect sizes, and $C(z) = \phi(z)/\Phi(z)$, where $z = (\mu -$

$\eta)/\sigma$, and μ and σ is the mean and standard deviation of the post-test effect size distribution, and η is a cutoff above which effect sizes are selected for follow-up. Lastly, $\phi()$ is the probability density function and $\Phi()$ is the cumulative distribution function of the normal distribution.

This formula assumes that both post-test and follow-up effect sizes are normally distributed with the same mean and variance (Khan & Olivier, 2018). These are in general strong assumptions but in our setting, they correspond relatively well to the assumption that RTM is the only fade-out mechanism. That is, the assumptions provide a scenario where we can examine if it seems likely that RTM can explain all fade-out. If the RTM effect is smaller than the fade-out we observe also under these assumptions, then it seems unlikely to be the only fade-out mechanism.

To use the formula, we needed estimates of the mean, the variance and the standard deviation, and the post-test to follow-up correlation. Because the majority of our studies were included in Dietrichson et al., (2021), we used the main meta-analytic sample in that review for the calculation: $\sigma^2 = 0.0996$, $\rho_{pf} = \tau^2/\sigma^2 = 0.673$ where τ^2 is the between-study variance (i.e., $\tau^2 = 0.067$), and $\mu = 0.30$, which is equal to the post-test weighted average effect size (see p. 30 in Dietrichson et al., 2021).

In the formula, η represents a strict cutoff. Because our data included studies and treatments with on average negative effect sizes and in general included a wide range of post-test effect sizes, it seems unlikely that there is one strict cutoff for when an intervention is followed up or not. We therefore used lower and upper bounds on η's. As our lower bound, we used $\eta = 0$. As our upper bound, we used the weighted average of the post-test effect sizes in the included studies, which was 0.38. In a sense, this is the empirical maximum in our sample because if

researchers on average used higher cutoffs, then we should have observed a higher weighted average (of course, they may very well have used a lower cutoff, something which seems likely given the empirical distribution we observe).

The calculations then yielded a lower bound for the RTM effect at 0.03 and an upper bound of 0.10. These effects correspond to 21% and 67% of the weighted average fade-out in the full sample, which was -0.15 (the full sample should be closest to the sample in Dietrichson et al., 2021). That is, also according to the upper bound on *k*, RTM explained far from all fade-out. It is thus not likely to be the only fade-out mechanism.

Supplemental Material F: Test of Assumption for Increasing Variance Mechanism

The increasing variance mechanism explains fade-out by the variance, and standard deviations used to calculate standardized mean differences, increasing between the post and follow-up test. A necessary condition for this mechanism is that the standard deviation is larger at follow-up than at post-test. This section provides a test of this assumption.

The test should preferably separate between reasons why the standard deviation may increase. The increasing variance mechanism causes fade-out when students' test scores become more heterogeneous over time. Another reason for the standard deviation to increase could be changes of the type of tests used. For example, older students may be tested using tests with more test items or a broader content, which seems likely to increase the standard deviation but may also increase the intervention effects to a similar degree.

A motivating example is the following: Assume 10 treatment and 10 control group students. Receiving the intervention implies that all treatment group students get all items correct on both tests, whereas the control group students get 50% right (i.e., the effect and heterogeneity is constant). Let the number of items be 10 at post-test and 20 at follow-up. Then the standard deviation is around 2.565 at post-test and 5.130 at follow-up, which makes, e.g., Cohen's *d* equal at both occasions (1.9). The equality of effect sizes seems right because there is no fade-out, the effect on skills is the same at both time points.

However, just testing whether the standard deviation is larger at follow-up using this type of test would, with enough statistical power, find that it is. Because it is the relation between the test score and the standard deviation that matters, we need a test that does not confound increases due to student heterogeneity with increases in the number of test items or content. This reasoning

suggests that we cannot include all our effect sizes but we could use the subset of tests with identical test items, for which neither content nor the number of items change.

We therefore used subset of effect sizes based on identical test items (but do not require the raw scores as in the tests of H5a and b), which yielded 20 studies and 148 effect sizes. We used the following two specifications:

$$ln\left(\frac{SD_{ks}^{fp}}{SD_{ks}^{post}}\right) = \beta_0 I_{1-11} + \beta_1 I_{\geq 12} + \mu_s + \upsilon_{ks} + \varepsilon_{ks}$$

$$ln\left(\frac{SD_{ks}^{fp}}{SD_{ks}^{post}}\right) = \beta_2 Duration + \beta_3 Duration^2 + \mu_s + \upsilon_{ks} + \varepsilon_{ks}$$

where $ln()$ is the natural logarithm, SD_{ks}^{fp} is the standard deviation of effect size *k* in study *s* at follow-up, SD_{ks}^{post} is the corresponding standard deviation of the anchor of *k* at post-test, I_{1-11} is an indicator for effect sizes with a duration between post and follow-up between 1 and 11 months, $I_{\geq 12}$ is an indicator for durations over 12 months, *Duration* is a continuous variable measuring the post to follow-up duration in months, *Duration*^{2} is the square of *Duration*, and μ_s, υ_{ks}, and ε_{ks} are defined as in the main text. Because there was only one study with effect sizes based on identical test items and post to follow-up duration over 17 months, we could not include an indicator for durations over 17 months (as in the primary analysis).

We estimated whether β_0 and β_1 were different from zero using the same CHE-RVE model and weighting scheme as in our primary analysis. We found neither large nor significant differences between the post and follow-up standard deviations for durations between 1-11 and over 12 months ($\hat{\beta}_0$ = -0.018, 95% CI = [-0.07, 0.03]), $\hat{\beta}_1$ = 0.004, 95% CI = [-0.21, 0.22]), and, similarly, the coefficients on *Duration* and *Duration*^{2} were not significantly different from zero

($\hat{\beta}_2$ = -0.04, 95% CI = [-0.15, 0.07]), $\hat{\beta}_3$ = 0.003, 95% CI = [-0.01, 0.02])). Although these results did not support the increasing variance mechanism, we could only use a subset of our studies and none of the studies included in this subset had a longer duration than 24 months. Thus, increasing variance may still be an important mechanism for longer follow-ups.

Supplemental Material G: Assumption on Trends in Relation to Outside Sample

The tests of H6a and H6b assume that the outside samples, on average, would have developed in the same pace as the treatment (H6a) and the control groups (H6b), had there not been an intervention. This assumption is similar to the parallel trends-assumption needed to identify effects in difference-in-differences analyses (e.g., Abadie, 2005). As in such analyses, an indirect test of the assumption would be to examine the pre-intervention trends. Because there is only one pre-test occasion in all our included studies, examining pre-trends would only be meaningful for the control group. However, for two reasons, the results of such an examination would be difficult to interpret. First, several studies select participants based on them scoring below a cutoff on one or more pre-tests. This selection ought to produce regression to the mean and the control group would therefore be expected to improve more than the outside sample between the pre and post-test. Second, pre and post-tests are typically conducted at different times of the year, which may make the comparison to a norm-referencing sample problematic if tests for this sample are not conducted with a similar interval.

Instead, we probed the sensitivity of our estimates using estimates of the trends from the earlier literature. As mentioned in the main text, the evidence of how at-risk students develop compared to other student groups is mixed. Studies have found increasing, decreasing, or no change of the gaps across different school periods (e.g., Caro et al., 2009; Geary et al., 2012; Halle et al., 2012; Heckman, 2006; Jordan et al., 2003; Lipsey et al., 2012; Pace et al., 2019; Sirin, 2005; von Hippel et al., 2018). Importantly, the typical increase or decrease seems small over time periods similar to the duration between post and follow-up tests in our analysis (median was 7 and the weighted average 10.7 months). We illustrate how increases and decreases

would affect our estimates below, and then report the findings of each cited study in some more detail.

To recapitulate, we used the following specification in the test of H6a and H6b:

$$ES_{ks}^{norm} = \alpha + \beta FU + \mu_s + \upsilon_{ks} + \varepsilon_{ks}$$

where FU is an indicator for follow-up tests. Because ES_{ks}^{norm} denotes the (standardized) difference between, e.g., the treatment group and the outside sample, at post and follow-up tests, $\hat{\beta}$ estimates the change between post and follow-up tests. That is, it is similar to the coefficient on a treatment indicator in a 2-group x 2-period difference-in-differences design.

First, note that a trend of decreasing differences between the treatment groups and outside samples would imply that our negative estimates in the primary analysis represent a break from the trend. In turn, estimates that took the trend into account would be more negative. Vice versa with increasing differences, estimates that took an increasing trend into account would be less negative (or turn positive).

To exemplify how such violations of the assumption of parallel trends would affect our estimates, we use data from Lipsey et al. (2012; see in particular p. 29-31). We exemplify with this study because it reported differences in metrics and for outcome measures that are relatively easy to compare to our estimates, and did so for a representative sample of groups that were similar to our sample. Lipsey et al. (2012) reported differences in reading and math between free- and reduced-price lunch (FRPL) eligible and ineligible students as well as minority and majority students in grade 4, 8, and 12 from the National Assessment of Educational Progress (NAEP). This test is of a similar type to the ones we included and the NAEP is representative of the US student body, where a large majority of our interventions were conducted. FRPL eligible and

minority students are overrepresented in many of our samples, so should be reasonably close to our treatment groups whereas FRPL ineligible and majority students resemble our outside samples.

The gap between FRPL eligible and ineligible students decreased over both periods and in both subjects. Focusing on grades 4 and 8, which are closest to the studies in the primary sample used to test H6a and b, the effect size difference in reading (math) is -0.74 (-0.85) in grade 4, and -0.66 (-0.80) in grade 8. The gap between Hispanics and Whites also decreased, as did the Black-White reading difference (both decreased less than the FRPL gap). The Black-White math difference however increased from -0.99 in grade 4 to -1.04 in grade 8.

Importantly though, assuming a constant rate of change across grades, the magnitudes would be small over time periods equal to the duration between post and follow-up in our sample, which were slightly shorter than a grade-to-grade change. The largest grade-to-grade decrease in the NAEP was -0.02 (-0.08 divided by 4 grades), which, when added to our estimates in the primary analysis would yield approximately -0.14 and -0.07. The largest increase was 0.0125, which, if considered in our estimations, would yield estimates of around -0.1075 and -0.0375. That is, there would be no substantial change to the estimates and conclusions.

Lipsey et al. (2012) also reported similar comparisons using results on the Stanford Achievement Test (SAT 9) from a large urban district. In this less representative sample, the swings were larger but again, there were both decreases and increases. The FRPL-gap decreased in both subjects, the Black-White gap decreased in reading and increased in math, whereas the Hispanic-White gap increased in both subjects. The largest change was the increase in the Hispanic-White gap in math, which was 0.09 grade-to-grade. All other grade-to-grade changes

were below 0.03. That is, out of the 12 comparisons (3 groups x 2 subjects x 2 tests) reported in Lipsey et al. (2012), only one would result in substantial changes to our results. Note also that 8 out 12 comparisons showed decreasing gaps, which would strengthen our conclusions regarding the treatment-induced (t) and treatment-induced (c) mechanisms.

Regarding the other studies, three reported changes in the gaps between groups of students defined by SES-related indicators. The changes in the gap between Canadian high and low SES students reported in Caro et al. (2009) seem very small up to age 11 (see e.g., Figure 1, p. 573), whereas the gap widens between ages 11 to 15. Von Hippel et al. (2018) used the ECLS-K:1999 and ECLS-K:2011 and found very small changes of the gap between the quintiles of SES between kindergarten and grade 2 (see e.g., Figure 2, p. 335). Heckman (2006) showed a slight divergence in math from age 6 to 12, measured as percentile ranks, between groups of students defined by the two lowest and highest income quintiles (Figure 1, p. 1900). Both Von Hippel et al. (2018) and Heckman (2006) concluded that the gaps mainly appeared before children enroll in schools.

Other studies defined at-risk groups in terms of earlier performance on skills test. Halle et al. (2012) found mainly decreasing differences in both math and reading between different definitions of at-risk groups in their analyses of ECLS-K and NICHD SECCYD data (see in particular Figure 1, 2, 4, and 5 p. 49-52). Pace et al. (2019) also used the NICHD SECCYD data to examine how early math, reading, language, and social-emotional skills predict the growth of the same skills. With the partial exception of language skills, higher levels of skills mainly predicted slower growth in math and reading both between grade 1 and 3, and between grade 3 and 5. Jordan et al. (2003) reported both increases and decreases in the gap between students

with poor and good “math fact mastery” from fall of second grade to spring of third grade on standardized tests (see Figure 2, p. 110). The modal change in these three studies was again small.

A partial exception to the pattern of small changes was reported by Geary et al. (2012). They found both convergence and divergence between typically achieving students (TA), low achieving students (LA), and students with mathematical learning difficulties (MLD). Some of the grade-to-grade changes were small, but there were also examples of changes that seem quite substantial (see e.g., Figure 1, p. 213). One potential explanation is the small sample sizes, particularly in the LA (n = 29) and MLD (n = 16) groups (in the TA group, n = 132).

Sirin (2005) conducted a large-scale meta-analysis of the correlation between students’ SES and achievement. Table 4, p. 435, shows the correlation broken down into categories defined by grade level. The correlation is lowest in kindergarten (0.19) and highest in middle school (0.31) with elementary school (0.27) and high school (0.26) in between. Although an increase in the correlation across grades may be caused by diverging trends, there was also, in that case, convergence between middle and high school, and a small change between elementary and middle school.

Taken together, it seems unclear from these studies whether we should expect increasing or decreasing differences between treatment and control groups and outside samples. The findings were quite mixed, also within studies. However, a large majority of the reported changes were small, which is reassuring as the impact on our estimates would likely then also be small. As mentioned in the main text, the analysis of intervention effects in Protzko (2015), which was based on normed IQ-tests, showed virtually no change in the average relation between the

control group and the norm-referenced sample. Protzko's findings are thus in line with the main message from the studies we surveyed here.

References Supplemental Material

References marked with an asterisk indicate records included in the meta-analysis.

Abadie, A. (2005). Semiparametric difference-in-differences estimators. Review of Economic Studies, *72*(1), 1-19. https://doi.org/10.1111/0034-6527.00321

*Alan, S., Boneva, T., & Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *Quarterly Journal of Economics*, *134*(3), 1121–1162. https://doi.org/10.1093/qje/qjz006

Bailey, D. H., Duncan, G. J., Watts, T., Clements, D. H., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *American Psychologist*, *73*(1), 81–94. https://doi.org/10.1037/amp0000146

*Bailey, D. H., Fuchs, L. S., Gilbert, J. K., Geary, D. C., & Fuchs, D. (2020). Prevention: Necessary but insufficient? A 2-year follow-up of an effective first-grade mathematics intervention. *Child Development*, *91*(2), 382–400. https://doi.org/10.1111/cdev.13175

Barnett, A. G., Van Der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

*Blachman, B. A., Schatschneider, C., Fletcher, J. M., Francis, D. J., Clonan, S. M., Shaywitz, B. A., & Shaywitz, S. E. (2004). Effects of intensive reading remediation for second and third graders and a 1-year follow-up. *Journal of Educational Psychology*, *96*(3), 444–461. https://doi.org/10.1037/0022-0663.96.3.444

*Blachman, B. A., Schatschneider, C., Fletcher, J. M., Murray, M. S., Munger, K. A., & Vaughn, M. G. (2014). Intensive reading remediation in grade 2 or 3: Are there effects a decade

later? *Journal of Educational Psychology*, *106*(1), 46–57. https://doi.org/10.1037/a0033663

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

Caro, D. H., McDonald, J. T., & Willms, J. D. (2009). Socio-economic status and academic achievement trajectories from childhood to adolescence. *Canadian Journal of Education/Revue Canadienne de l'éducation*, *32*(3), 558–590.

*Cirino, P. T., Vaughn, S., Linan-Thompson, S., Cardenas-Hagan, E., Fletcher, J. M., & Francis, D. J. (2009). One-year follow-up outcomes of Spanish and English interventions for English language learners at risk for reading problems. *American Educational Research Journal*, *46*(3), 744–781. https://doi.org/10.3102/0002831208330214

*Clarke, B., Doabler, C., Smolkowski, K., Kurtz Nelson, E., 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*(4), 607–634. https://doi.org/10.1080/19345747.2015.1116034

*Clarke, B., Doabler, C. T., Sutherland, M., Kosty, D., Turtura, J., & Smolkowski, K. (2023). Examining the impact of a first grade whole number intervention by group size. *Journal of Research on Educational Effectiveness*, *16*(2), 326–349. https://doi.org/10.1080/19345747.2022.2093299

*Cleary, T. J., Velardi, B., & Schnaidman, B. (2017). Effects of the self-regulation empowerment program (SREP) on middle school students' strategic skills, self-efficacy, and

mathematics achievement. *Journal of School Psychology*, *64*, 28–42. https://doi.org/10.1016/j.jsp.2017.04.004

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research*, *87*(2), 243–282. https://doi.org/10.3102/0034654316687036

Dietrichson, J., Filges, T., Seerup, J. K., Klokker, R. H., Viinholt, B. C. A., Bøg, M., & Eiberg, M. (2021). Targeted school-based interventions for improving reading and mathematics for students with or at risk of academic difficulties in Grades K-6: A systematic review. *Campbell Systematic Reviews*, *17*(2), e1152. https://doi.org/10.1002/cl2.1152

*Doabler, C. T., Clarke, B., Turtura, J. E., Sutherland, M., Gersib, J. A., Lesner, T., ... & Kosty, D. (2024). Proficiency with number concepts and operations: Replicating the efficacy of a first-grade mathematics intervention. *Journal of Learning Disabilities*, 57(4), 224–241. https://doi.org/10.1177/00222194231209017

*Dyson, N. I., Jordan, N. C., & Glutting, J. (2013). A number sense intervention for low-income kindergartners at risk for mathematics difficulties. *Journal of Learning Disabilities*, *46*(2), 166–181. https://doi.org/10.1177/0022219411410233

*Dyson, N. I., Jordan, N. C., Beliakoff, A., & Hassinger-Das, B. (2015). A kindergarten number-sense intervention with contrasting practice conditions for low-achieving children. *Journal for Research in Mathematics Education*, *46*(3), 331–370. https://doi.org/10.5951/jresematheduc.46.3.0331

*Dyson, N. I., Jordan, N. C., Rodrigues, J., Barbieri, C., & Rinne, L. (2020). A fraction sense intervention for sixth graders with or at risk for mathematics difficulties. *Remedial and Special Education*, *41*(4), 244–254. https://doi.org/10.1177/0741932518807139

*Elbro, C., & Petersen, D. K. (2004). Long-term effects of phoneme awareness and letter sound training: An intervention study with children at risk for dyslexia. *Journal of Educational Psychology*, *96*(4), 660–670. https://doi.org/10.1037/0022-0663.96.4.660

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *Robumeta: Robust variance meta-regression*. Retrieved from https://github.com/zackfisher/robumeta

*Fuchs, L. S., Fuchs, D., Sterba, S. K., Barnes, M. A., Seethaler, P. M., & Changas, P. (2022). Building word-problem solving and working memory capacity: A randomized controlled trial comparing three intervention approaches. *Journal of Educational Psychology*, *114*(7), 1633–1653. https://doi.org/10.1037/edu0000752

*Fälth, L., Gustafson, S., Tjus, T., Heimann, M., & Svensson, I. (2013). Computer-assisted interventions targeting reading skills of children with reading disabilities–A longitudinal study. *Dyslexia*, *19*(1), 37–53. https://doi.org/10.1002/dys.1450

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

*Goodson, B., Wolf, A., Bell, S., Turner, H., & Finney, P. B. (2010). The effectiveness of a program to accelerate vocabulary development in kindergarten (VOCAB): Kindergarten

final evaluation report. NCEE 2010-4014. *National Center for Education Evaluation and Regional Assistance*.

*Goodson, B., Wolf, A., Bell, S., Turner, H., & Finney, P. B. (2011). Effectiveness of a program to accelerate vocabulary development in kindergarten (VOCAB): First grade follow-up impact report and exploratory analyses of kindergarten impacts. Final report. NCEE 2012-4009. *National Center for Education Evaluation and Regional Assistance*.

*Gunn, B., Smolkowski, K., Biglan, A., Black, C., & Blair, J. (2005). Fostering the development of reading skill through supplemental instruction: Results for Hispanic and non-Hispanic students. *Journal of Special Education*, *39*(2), 66–85. https://doi.org/10.1177/00224669050390020301

*Gustafson, S., Fälth, L., Svensson, I., Tjus, T., & Heimann, M. (2011). Effects of three interventions on the reading skills of children with reading disabilities in grade 2. *Journal of Learning Disabilities*, *44*(2), 123–135. https://doi.org/10.1177/0022219410391187

Halle, T. G., Hair, E. C., Buchinal, M., Anderson, R., & Zaslow, M. (2012). *In the running for successful outcomes: Exploring the evidence for thresholds of school readiness*. Technical report, U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation, Washington D. C. Retrieved from http://aspe.hhs.gov/reports/running-successful-outcomes-exploring-evidence-thresholds-school-readiness-technical-report-0.

*Hallstedt, M. H., Klingberg, T., & Ghaderi, A. (2018). Short and long-term effects of a mathematics tablet intervention for low performing second graders. *Journal of Educational Psychology*, *110*(8), 1127–1148. https://doi.org/10.1037/edu0000264

*Hansson, Å. (2015). *Effekter av intensivundervisning i matematik. Utvärdering av ett pilotprojekt med personlig tränare i matematik för elever i behov av särskilt stöd* (RIPS: Rapporter Från Institutionen För Pedagogik Och Specialpedagogik No. nr 7). Institutionen för pedagogik och specialpedagogik.

*Hatz, H., & Sachse, S. (2010). Prävention von lese-rechtschreibstörungen: Auswirkungen eines trainings phonologischer bwusstheit und eines rechtschreibutrainings im ersten schuljahr auf den schriftspracherwerb bei risikokindern. *Zeitschrift für Entwicklungspsychologie Und Pädagogische Psychologie*, *42*(4), 226–240. https://doi.org/10.1026/0049-8637/a000021

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*(5782), 1900–1902. https://doi.org/10.1126/science.1128898

*Hurry, J., & Sylva, K. (2007). Long-term outcomes of early reading intervention. *Journal of Research in Reading*, *30*(3), 227–248. https://doi.org/10.1111/j.1467-9817.2007.00338.x

*Johnson, H., McNally, S., Rolfe, H., Ruiz-Valenzuela, J., Savage, R., Vousden, J., & Wood, C. (2019). Teaching assistants, computers and classroom management. *Labour Economics*, *58*, 21–36. https://doi.org/10.1016/j.labeco.2019.02.006

*Jones, C. J., & Christian, M. (2021). The results of a randomized control trial evaluation of the SPARK literacy program: An innovative approach that pairs one-on-one tutoring with family engagement. *Journal of Education for Students Placed at Risk*, *26*(3), 185–209. https://doi.org/10.1080/10824669.2020.1809419

*Jones, C. J., Reeves, M., Li, D., & Gilman, L. (2023). What is the sustained impact of Future Forward on reading achievement, attendance, and special education placement 5 years

after participation? *Educational Evaluation and Policy Analysis*, online advance access. https://doi.org/10.3102/016237372311826

*Jordan, N. C., Glutting, J., Dyson, N., Hassinger-Das, B., & Irwin, C. (2012). Building kindergartners' number sense: A randomized controlled study. *Journal of Educational Psychology*, *104*(3), 647–660. https://doi.org/10.1037/a0029018

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). Arithmetic fact mastery in young children: A longitudinal investigation. *Journal of Experimental Child Psychology*, *85*(2), 103–119. https://doi.org/10.1016/S0022-0965(03)00032-8

Khan, M., & Olivier, J. (2018). Quantifying the regression to the mean effect in Poisson processes. *Statistics in Medicine*, *37*(26), 3832–3848. https://doi.org/10.1002/sim.7850

*Konerding, M., Bergström, K., Lachmann, T., & Klatte, M. (2020). Effects of computerized grapho-phonological training on literacy acquisition and vocabulary knowledge in children with an immigrant background learning German as L2. *Journal of Cultural Cognitive Science*, *4*, 368–383. https://doi.org/10.1007/s41809-020-00064-3

Linden, A. (2013). Assessing regression to the mean effects in health care initiatives. *BMC Medical Research Methodology*, *13*(1), 1–7. https://doi.org/10.1186/1471-2288-13-119

*Lindström-Sandahl, H., Danielsson, H., Elwér, Å., Samuelsson, S., & Samuelsson, J. (2024). A 1-year follow-up of two intensive supplemental reading and mathematics programs in Swedish elementary school. In Lindström-Sandahl, H. *Early elementary school interventions in reading and mathematics* (Doctoral dissertation, Linköping University Electronic Press). Linköping University. https://doi.org/10.3384/9789180754750

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *National Center for Special Education Research*.

*Lopez-Pedersen, A., Mononen, R., Aunio, P., Scherer, R., & Melby-Lervåg, M. (2023). Improving numeracy skills in first graders with low performance in early numeracy: A randomized controlled trial. *Remedial and Special Education*, *44*(2), 126–136. https://doi.org/10.1177/07419325221102537

Marston, D. (1996). A comparison of inclusion only, pull-out only, and combined service models for students with mild disabilities. *Journal of Special Education*, *30*(2), 121–132. https://doi.org/10.1177/002246699603000

Pace, A., Alper, R., Burchinal, M. R., Golinkoff, R. M., & Hirsh-Pasek, K. (2019). Measuring success: Within and cross-domain predictors of academic and social trajectories in elementary school. *Early Childhood Research Quarterly*, *46*, 112–125. https://doi.org/10.1016/j.ecresq.2018.04.001

*Powell, S. R., Berry, K. A., Fall, A.-M., Roberts, G., Fuchs, L. S., & Barnes, M. A. (2021). Alternative paths to improved word-problem performance: An advantage for embedding prealgebraic reasoning instruction within word-problem intervention. *Journal of Educational Psychology*, *133*(5), 898–910. https://doi.org.10.1037/edu0000513

*Powell, S. R., Berry, K. A., Fall, A. M., Roberts, G., Barnes, M. A., Fuchs, L. S., ... & Lin, X. (2022). Does word-problem performance maintain? Follow-up one year after implementation of a word-problem intervention. *Journal of Research on Educational Effectiveness*, *15*(1), 52–77. https://doi.org/10.1080/19345747.2021.1961332

*Prediger, S., & Wessel, L. (2018). Brauchen mehrsprachige jugendliche eine andere fach-und sprachintegrierte förderung als einsprachige? *Zeitschrift für Erziehungswissenschaft*, *21*(2), 361–382. https://doi.org/10.1007/s11618-017-0785-8

Protzko, J. (2015). The environment in raising early intelligence: A meta-analysis of the fadeout effect. *Intelligence*, *53*, 202–210. https://doi.org/10.1016/j.intell.2015.10.006

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, *10*(1), 57–71. https://doi.org/10.1002/jrsm.1332

*Ryder, J. F., Tunmer, W. E., & Greaney, K. T. (2008). Explicit instruction in phonemic awareness and phonemically based decoding skills as an intervention strategy for struggling readers in whole language classrooms. *Reading and Writing*, *21*(4), 349–369. https://doi.org/10.1007/s11145-007-9080-z

*Schanzenbach, D. W. (2006). What have researchers learned from Project STAR? *Brookings Papers on Education Policy*, *9*, 205–228. https://doi.org/http://www.jstor.org/stable/20067282

*Schüler-Meyer, A., Prediger, S., Kuzu, T., Wessel, L., & Redder, A. (2019). Is formal language proficiency in the home language required to profit from a bilingual teaching intervention in mathematics? A mixed methods study on fostering multilingual students' conceptual understanding. *International Journal of Science and Mathematics Education*, *17*(2), 317–339. https://doi.org/10.1007/s10763-017-9857-8

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453.

*Somers, M.-A., Corrin, W., Sepanik, S., Salinger, T., Levin, J., & Zmach, C. (2010). *The enhanced reading opportunities study. Final report: The impact of supplemental literacy courses for struggling ninth-grade readers* (NCEE 2010-4021). National Center for Education Evaluation and Regional Assistance.

*Svensson, I., Nordström, T., Lindeblad, E., Gustafson, S., Björn, M., Sand, C., ... & Nilsson, S. (2021). Effects of assistive technology for students with reading and writing disabilities. *Disability and Rehabilitation: Assistive Technology*, *16*(2), 196-208. https://doi.org/10.1080/17483107.2019.1646821

*Tolan, P., Gorman-Smith, D., & Henry, D. (2004). Supporting families in a high-risk setting: Proximal effects of the SAFEChildren preventive intervention. *Journal of Consulting and Clinical Psychology*, *72*(5), 855–869. https://doi.org/10.1037/0022-006X.72.5.855

*Toll, S. W., & Van Luit, J. E. (2014). Effects of remedial numeracy instruction throughout kindergarten starting at different ages: Evidence from a large-scale longitudinal study. *Learning and Instruction*, *33*, 39–49. https://doi.org/10.1016/j.learninstruc.2014.03.003

*Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Herron, J., & Lindamood, P. (2010). Computer-assisted instruction to prevent early reading difficulties in students at risk for dyslexia: Outcomes from two instructional approaches. *Annals of Dyslexia*, *60*(1), 40–56. https://doi.org/10.1007/s11881-009-0032-y

*Torgesen, J., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., ... Haan, C. (2006). National assessment of Title I: Interim report. Volume II: Closing the reading gap: First year findings from a randomized trial of four reading interventions for striving

readers. NCEE 2006-4002. *National Center for Education Evaluation and Regional Assistance*.

*Torgesen, J., Schirm, A., Castner, L., Vartivarian, S., Mansfield, W., Myers, D., ... Haan, C. (2007). National assessment of Title I. Final report. Volume II: Closing the reading gap–findings from a randomized trial of four reading interventions for striving readers. NCEE 2008-4013. *National Center for Education Evaluation and Regional Assistance*.

*Vadasy, P. F., Jenkins, J. R., & Pool, K. (2000). Effects of tutoring in phonological and early reading skills on students at risk for reading disabilities. *Journal of Learning Disabilities*, *33*(6), 579–590. https://doi.org/10.1177/002221940003300606

*Vadasy, P. F., & Sanders, E. A. (2010). Efficacy of supplemental phonics-based instruction for low-skilled kindergarteners in the context of language minority status and classroom phonics instruction. *Journal of Educational Psychology*, *102*(4), 786–803. https://doi.org/10.1037/a0019639

*Vadasy, P. F., & Sanders, E. A. (2011). Efficacy of supplemental phonics-based instruction for low-skilled first graders: How language minority status and pretest characteristics moderate treatment response. *Scientific Studies of Reading*, *15*(6), 471–497. https://doi.org/10.1080/10888438.2010.501091

*Vadasy, P. F., & Sanders, E. A. (2012). Two-year follow-up of a kindergarten phonics intervention for English learners and native English speakers: Contextualizing treatment impacts by classroom literacy instruction. *Journal of Educational Psychology*, *104*(4), 987–1005. https://doi.org/10.1037/a0028163

*Vadasy, P. F., & Sanders, E. A. (2013). Two-year follow-up of a code-oriented intervention for lower-skilled first-graders: The influence of language status and word reading skills on third-grade literacy outcomes. *Reading and Writing*, *26*(6), 821–843. https://doi.org/10.1007/s11145-012-9393-4

*Van De Rijt, B. A., & Van Luit, J. E. (1998). Effectiveness of the additional early mathematics program for teaching children early mathematics. *Instructional Science*, *26*(5), 337–358. https://doi.org/10.1023/A:1003180411209

*Vaughn, S., Cirino, P. T., Linan-Thompson, S., Mathes, P. G., Carlson, C. D., Hagan, E. C., ... Francis, D. J. (2006). Effectiveness of a Spanish intervention and an English intervention for English-language learners at risk for reading problems. *American Educational Research Journal*, *43*(3), 449–487. https://doi.org/10.3102/00028312043003449

*Vaughn, S., Cirino, P. T., Tolar, T., Fletcher, J. M., Cardenas-Hagan, E., Carlson, C. D., & Francis, D. J. (2008). Long-term follow-up of Spanish and English interventions for first-grade English language learners at risk for reading problems. *Journal of Research on Educational Effectiveness*, *1*(3), 179–214. https://doi.org/10.1080/19345740802114749

*Vaughn, S., Linan-Thompson, S., Mathes, P. G., Cirino, P. T., Carlson, C. D., Pollard-Durodola, S. D., ... Francis, D. J. (2006). Effectiveness of Spanish intervention for first-grade English language learners at risk for reading difficulties. *Journal of Learning Disabilities*, *39*(1), 56–73. https://doi.org/10.1177/00222194060390010601

*Vaughn, S., Mathes, P., Linan-Thompson, S., Cirino, P., Carlson, C., Pollard-Durodola, S., ... Francis, D. (2006). Effectiveness of an English intervention for first-grade English

language learners at risk for reading problems. *The Elementary School Journal*, *107*(2), 153–180. https://doi.org/10.1086/510653

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in reading and math skills forms mainly before kindergarten: A replication, and partial correction, of "Are schools the great equalizer?". *Sociology of Education*, *91*(4), 323–357. https://doi.org/10.1177/0038040718801760

*Wolff, U. (2011). Effects of a randomised reading intervention study: An application of structural equation modelling. *Dyslexia*, *17*(4), 295–311. https://doi.org/10.1002/dys.438

*Wolff, U. (2016). Effects of a randomized reading intervention study aimed at 9-year-olds: A 5-year follow-up. *Dyslexia*, *22*(2), 85–100. https://doi.org/10.1002/dys.1529