Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1069

# Using Meta-Analytic Data to Examine Fadeout and Persistence of Intervention Impacts on Constrained and Unconstrained Skills

Mindy L. Rosengarten
Teachers College, Columbia University

Emma R. Hart
Teachers College, Columbia University

Drew H. Bailey
University of California- Irvine

Meghan P. McCormick
Overdeck Family Foundation

Benjamin J. Lovett
Teachers College, Columbia University

Tyler W. Watts
Teachers College, Columbia University

Recent reviews of the educational intervention literature have noted patterns of intervention impact fadeout on cognitive skills, whereby skill trajectories between children in the intervention and control group converge in the years following the end of the intervention. Some early childhood education (ECE) researchers have suggested that skill type, specifically whether a skill is "constrained" or "unconstrained" may explain variation in fadeout trajectories. The Constrained Skills View proposes that unconstrained skills, which are thought to develop across the life course, may show more persistent impacts than constrained skills, which are eventually mastered by all. For a broad, short-term test of this theory, we used the Meta-Analysis of Educational RCTs with Follow-up (MERF) to examine trajectories of fadeout and persistence by skill type across a variety of educational interventions tested in childhood and adolescence. The majority of impacts in our sample (91%) were on measures of reading and language skills. We modeled patterns of intervention impact persistence and fadeout six to twelve months after the interventions ended. After coding outcomes as "constrained" or "unconstrained," we found no evidence that impacts on unconstrained skills persisted more than impacts on constrained skills. Rather, in some model specifications, impacts on constrained skills showed slightly more short-term persistence than impacts on unconstrained skills.

VERSION: October 2024

Suggested citation: Rosengarten, Mindy L., Emma R. Hart, Drew H. Bailey, Meghan P. McCormick, Benjamin J. Lovett, and Tyler W. Watts. (2024). Using Meta-Analytic Data to Examine Fadeout and Persistence of Intervention Impacts on Constrained and Unconstrained Skills. (EdWorkingPaper: 24-1069). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/qn5q-ms33

# Using Meta-Analytic Data to Examine Fadeout and Persistence of Intervention Impacts on Constrained and Unconstrained Skills

Mindy L. Rosengarten[1], Emma R. Hart[1], Drew H. Bailey[2], Meghan P. McCormick[3], Benjamin J. Lovett[1], and Tyler W. Watts[1*]

1. Teachers College, Columbia University
2. University of California- Irvine
3. Overdeck Family Foundation

## Author Note

We have no conflicts of interest to disclose.

Correspondence regarding this article addressed to Tyler W. Watts, Teachers College, Columbia University, 463 Grace Dodge Hall, 525 W. 120 St. New York, NY 10025.

Email: tww2108@tc.columbia.edu

## Acknowledgments

Thank you to Christina Weiland for the support in theory development, coding measures of measures, and providing feedback on multiple iterations of this manuscript.

Time spent on this project was supported by the National Institute of Child Health and Human Development (1R01HD095930-01A1 to TW), the National Science Foundation (DGE-2036197 to EH), and the Institute of Education Sciences, U.S. Department of Education, (R305B200017 to Teachers College, Columbia University). We would like to thank the following research assistants whose efforts made this work possible (in alphabetical order): Helen Ding, Precious Elam, Simran Juneja, Susan Kruglinski, Gabby Lammano, Siyu Liang, Sha Luo, Opal Ofstedal, Fatmanur Ozay, Xinyu Pan, Spruha Reddy, John Schupbach, Maddie Scricco, Pritha Sengupta, Jessica Sperber, Devon Turner, Leo Weaver, and Josefa Westerman.

**Abstract**

Recent reviews of the educational intervention literature have noted patterns of intervention impact fadeout on cognitive skills, whereby skill trajectories between children in the intervention and control group converge in the years following the end of the intervention. Some early childhood education (ECE) researchers have suggested that skill type, specifically whether a skill is "constrained" or "unconstrained" may explain variation in fadeout trajectories. The Constrained Skills View proposes that unconstrained skills, which are thought to develop across the life course, may show more persistent impacts than constrained skills, which are eventually mastered by all. For a broad, short-term test of this theory, we used the Meta-Analysis of Educational RCTs with Follow-up (MERF) to examine trajectories of fadeout and persistence by skill type across a variety of educational interventions tested in childhood and adolescence. The majority of impacts in our sample (91%) were on measures of reading and language skills. We modeled patterns of intervention impact persistence and fadeout six to twelve months after the interventions ended. After coding outcomes as "constrained" or "unconstrained," we found no evidence that impacts on unconstrained skills persisted more than impacts on constrained skills. Rather, in some model specifications, impacts on constrained skills showed slightly more short-term persistence than impacts on unconstrained skills.

*Keywords*: Constrained Skills View, meta-analysis, educational RCTs, fadeout

**Public Significance Statement**

This study compared the impacts of educational interventions on unconstrained skills, like reading comprehension, and constrained skills, like phonics, at the end of the intervention and in the months following the intervention. Although theory suggests that improvements to unconstrained skills should persist more than improvements to constrained skills, we found that

gains on both skill types faded out in the months following the interventions, with some evidence that effects on unconstrained skills faded to a greater extent. Our findings suggest that intervention effect fadeout remains an important issue across various types of skills, necessitating the need for further innovation from researchers and intervention developers.

## Using Meta-Analytic Data to Examine Fadeout and Persistence of Intervention Impacts on Constrained and Unconstrained Skills

Recent reviews of the education intervention literature suggest that program impacts on cognitive skills often fade in the years following the end of the intervention (Abenavoli, 2019; Bailey et al., 2020; Li et al., 2020; Protzko, 2015). Theoretical work suggests that intervention impacts on certain classes of skills may be less susceptible to fadeout than others (Bailey et al., 2017; Paris, 2005; Snow & Matthews, 2016). "Unconstrained skills" have been posited to be one skill type for which fadeout may be less prevalent (Paris, 2005). Unconstrained skills are those that lack a ceiling and develop continuously across the life course (e.g., vocabulary), whereas constrained skills are those that nearly all children will master and for which performance reaches a ceiling (e.g., phonemic awareness). Researchers have argued that intervention impacts on constrained skills may fade quickly after an intervention ends because children in the treatment group reach a ceiling in performance and have limited opportunities for further growth. Simultaneously, children in the control group catch up, thus diminishing advantages for the treatment group as all children reach mastery (Bailey et al., 2020; McCormick & Mattera, 2022).

Although this theory has received attention in both literacy (e.g., Suggate, 2016) and early childhood education literature (e.g., McCormick et al., 2022), it has yet to be systematically tested in a broad dataset of educational interventions targeting a range of skills and ages. The current study leveraged a dataset of educational randomized controlled trials (RCTs) to examine trajectories of short-term fadeout for constrained and unconstrained skills. We categorized 159 unique measures of child skills gathered across 54 interventions as either constrained or unconstrained. The ages and skill foci of these interventions were broad, allowing us to examine the Constrained Skill View across a host of educational interventions. We focused our analyses

on 223 effects measured consistently at posttest and 6- to 12-month follow-up, enabling us to examine impact trajectories over one year following the intervention. In the next sections, we introduce the Constrained Skills View and discuss its influence on various subfields of education. We then examine how this idea has been used to explain intervention impacts on skill development before discussing our empirical approach.

## Literature Review

### *Constrained Skills View*

The "Constrained Skill View" was proposed by Paris (2005) to categorize the various skills involved in the development of reading achievement. Paris (2005) described constrained skills as having a relatively small number of elements that can be quickly mastered, with examples including alphabet knowledge and phonics. For most constrained skills, virtually all students reach mastery within a relatively limited range of time. The growth trajectory of these skills is thought to follow a sigmoid curve wherein growth begins slowly, becomes rapid as learning increases, and slows again once mastery is achieved. Although many constrained skills are likely to be essential for the development of future reading skills, Paris argued that constrained skills become less predictive of future reading abilities as children age because the variance in performance shrinks as all children reach mastery. Thus, Paris explained, gains in constrained skills are transient because they are universally acquired. In contrast, Paris proposed that unconstrained skills have a wider scope and more elements to learn, with vocabulary and comprehension being primary examples (Snow & Matthews, 2016). These skills develop across the lifespan, show variation that can expand with time, and importantly, are not universally mastered. Because between-person variation persists across time, these skills are thought to be highly *statistically* predictive of future reading ability.

However, Paris cautioned that such predictions were unlikely to be causal or clearly informative for interventions. Regarding interventions, Paris predicted that reading programs targeting constrained skills due to their apparent predictive validity would likely yield disappointing results in the long term, as these skills would be unlikely to transfer to broader reading capacities and would be quickly mastered by students in the control group ("temporary acceleration of mastered skills," p. 198-199). Conversely, Paris argued that interventions targeting unconstrained skills should be more likely to have long-run effects on broader reading abilities as they are related to a wider scope of skills.

### *Application to Literacy Programs*

While it is difficult to directly test whether unconstrained skills play a causal role in shaping broader reading and literacy, educational intervention evaluations with follow-up assessments can provide helpful insights regarding the persistence of impacts on skills measured over time. If intervention effects on unconstrained skills persist more than constrained skills, this may suggest that these skills are instrumental in the development of broader literacy skills. If experimental boosts to unconstrained skills do not persist, this may indicate that there are more complex causal pathways by which unconstrained skills come to shape broader reading capacities (e.g., through a more complex network of other skills), or that unconstrained skills may not be uniquely instrumental to the development of later reading capacities.

To date, a handful of studies have explored whether: 1) interventions that focus on promoting unconstrained skills find more persistent intervention effects, and 2) whether intervention impacts on unconstrained skills (regardless of focus) are more persistent than impacts on constrained skills. Among studies that have examined intervention foci, few have explicitly investigated whether targeting constrained versus unconstrained capacities generates

more persistent impacts. Suggate's (2016) meta-analysis of 71 (quasi-)experimental reading interventions run during Pre-K to grade six documented intervention content, measures, and methodology. It should be noted the Suggate (2016) meta-analysis is one of the eight used to source papers used in the current meta-analysis. Suggate found mixed evidence regarding the Constrained Skills View. Consistent with predictions, interventions targeting phonics produced effects that faded substantially between posttest and one-year follow-up (average follow-up time was about 11 months), and interventions targeting reading comprehension had persistent or growing effects between posttest and follow-up. However, phonemic awareness interventions, targeting perhaps the most constrained skillset in reading development, also produced rather persistent effects. Other meta-analyses similarly provide mixed evidence when applying the Constrained Skills View to intervention foci, with some reporting similar persistence for interventions regardless of skill focus (Silverman et al., 2020[1]) and others reporting greater fadeout for interventions targeting unconstrained skills compared to constrained skills (Fikrat-Wevers et al., 2021).[2]

The meta-analytic evidence is no clearer when examining whether impacts on constrained versus unconstrained skills, regardless of intervention focus, show different trajectories of skill fadeout. Suggate (2016) reported that although spelling skills (categorized as unconstrained) showed more persistent effects regardless of intervention focus, posttest impacts on prereading (categorized as constrained) and comprehension skills (categorized as unconstrained) showed some degree of fadeout at follow-up. Similarly, mixed findings were

[1] Note that Silverman et al. (2020) only report impacts on vocabulary, reading comprehension, and listening comprehension.

[2] Neither Fikrat-Wevers et al. (2021) nor Silverman et al. (2020) interpret their findings along the constrained/unconstrained skills continuum.

reported in a meta-analysis of (quasi-)experimental family literacy programs involving children from birth to six years of age, as authors found no support for the idea that intervention effects, regardless of focus, on unconstrained skills showed greater persistence (Fikrat-Wevers et al., 2021). Here, authors broadly characterized content and skills as either code-related (phonological awareness, letter knowledge; typically considered constrained) or reading comprehension-related (vocabulary; typically considered unconstrained). The largest posttest effects on children's skills were found for interventions focusing on code-related skills, which is unsurprising given that constrained skills are believed to be easier to change (Snow & Matthews, 2016). However, code-related skills (posttest effect of 0.48 SD that faded to 0.22 SD; 46% persistence) also showed less fadeout in the follow-up period (follow-up assessments ranged from six to 312 weeks after the intervention) compared to reading comprehension-related skills (posttest effect of 0.51 SD that faded to 0.09 SD; 18% persistence).

In summary, the extant literacy literature provides rather mixed evidence regarding the Constrained Skills View in terms of both intervention foci and skill development. However, the existing meta-analytic evidence is limited in several noteworthy ways. Meta-analytic studies do not often track impacts on the same skill over time, instead averaging skill impacts for interventions at post-test and follow-up time points (e.g., Suggate, 2016). This approach does not allow one to track whether impacts are maintained on the same skill over time, making it difficult to know if observed effects are due to persistence on a given skill, or due to transfer effects to other skills. Further, because researchers nonrandomly select studies and measures to be included at follow-up, averaging across all reported outcomes at a given timepoint may introduce bias due to researcher selective reporting (see review of this issue in Hart et al., in press). In the current paper, we attend to these issues with updated meta-analytic approaches to

tracking longitudinal effects that we describe below. Next, we discuss the broader applications of this theory to other educational research areas outside of literacy.

### *Broader Applications*

Though the Constrained Skills View was originally applied to studies of reading achievement, recent studies have expanded the idea to broader sets of academic skills, including mathematics achievement (McCormick et al., 2017; Spiegel et al., 2021). The Constrained Skills View has become popular in early education research as the field searches for skills that show persistence when boosted through education interventions. Bailey et al. (2017) argued that interventions targeting "trifecta skills," or skills that are (a) malleable, (b) crucial for academic success, and (c) unlikely to develop in the absence of the intervention, would produce longer-lasting impacts. When considering why patterns of intervention impacts might differ across constrained and unconstrained skills, it is not difficult to imagine why impacts on unconstrained skills might persist at higher rates. Unconstrained skills are believed to be less likely to receive direct instruction in the post-treatment period because they are harder to teach and assess (Snow & Matthew, 2016). As such, children in the control group should be slower to learn the skills acquired during the intervention.

The broad application of the Constrained Skills View beyond the realm of reading achievement parallels other theoretical skills categorizations in the psychological literature. For example, the "open" versus "closed" tasks dichotomy in the motor development and task performance literature reaches similar conclusions (Ackerman, 2007; Gu et al., 2019). Similar ideas have also been proposed in studies of children's mathematics achievement. For example, Rittle-Johnson and colleagues (2015) argued that procedural knowledge entails the ability to solve specific problem types, while conceptual knowledge requires the ability to reason abstractly about concepts and principles. As with constrained skills, procedural knowledge

includes fewer elements that can be quickly mastered, whereas conceptual knowledge can continue expanding through the life course. Similar to evidence on the role of unconstrained skills in shaping future literacy (Paris, 2005), children who demonstrate conceptual math knowledge early through the use of invented math strategies demonstrate a greater understanding of base-ten math concepts than children who rely on standard algorithms (Carpenter et al., 1998). Furthermore, children who utilize invented algorithms to solve math problems perform better on tasks that entail transferring knowledge to new problems (Carpenter et al., 1998). These findings are similar to claims of the importance of unconstrained skills in predicting future reading abilities (Paris, 2005).

**Fadeout in ECE and the Post-Intervention Environment**

Much of the recent research tracking intervention impact fadeout between constrained and unconstrained skills comes from evaluations of pre-k. In studies of pre-k effectiveness, measures of letter-word identification, number recognition, and print awareness have been cited as examples of constrained skills, while measures of mathematical reasoning, expressive and receptive vocabulary, and reading comprehension have been described as unconstrained skills (Johnson et al., 2022; McCormick et al., 2021; McCormick & Mattera, 2022; Whittingham et al., 2021). To our knowledge, only two RCTs have made a priori predictions of skill persistence based on the constrained/unconstrained skill classification model and documented such differences seven months to one year after the intervention (Grøver et al., 2024; Mattera et al., 2018; McCormick & Mattera, 2022). Other scholars (McCormick et al., 2021) have also interpreted results from existing experimental work using the constrained/unconstrained skills dichotomy.

There are more correlational studies that have reported consistent differences in skill persistence based on assessment classifications. These studies report that unconstrained skills show less fadeout compared to constrained skills into kindergarten (McCormick et al., 2021) and that Pre-K attenders outperform non-attenders on measures of unconstrained —but not constrained— skills in first grade (Ansari et al., 2023) and third grade (Johnson et al., 2022).

Evidence from the Pre-K literature suggests that when considering how the longer-term persistence patterns could differ based on the constrained/unconstrained skill dichotomy, the interplay between the skill type and the instructional environment is critical. Descriptive work has documented that kindergarten environments often focus much of their instruction on the promotion of "basic" skills, which likely reflects a heavy emphasis on "constrained" skills (Claessens et al., 2014). Indeed, a study in Boston suggests that the content of instruction in kindergarten can explain whether Pre-K effects are sustained, with more time spent in unconstrained instruction associated with the persistence of the Pre-K boost among Pre-K attendees (McCormick et al., 2022).

### *Challenges of Testing the Constrained Skills View*

Despite the potential strengths of the Constrained Skills View to explain trajectories of fadeout, there remain issues in its application. The categorization of skills as either constrained or unconstrained has been inconsistent, with variation from study to study. For example, while most studies utilizing this theory tend to focus on cognitive abilities, others have also described executive functioning and socioemotional skills as unconstrained skills (Ansari et al., 2023; Durkin et al., 2022). The question of what kinds of skills should be classified according to this continuum thus complicates its use. Many measures demand the application of both constrained and unconstrained skills. For example, Paris (2005) argued that the Comprehensive Test of

Phonological Processing (CTOPP) measures both constrained skills such as phonological memory, and unconstrained skills such as phonological awareness. In recent ECE applications of this theory, there are cases of some studies labeling a measure as constrained (McCormick et al., 2021) and a separate paper labeling a subscale of the same measure unconstrained (Johnson et al., 2022). Similarly, the field has yet to determine whether skills should be classified more generally versus granularly; for example, Barnett et al. (2018) referred to broad literacy and math as unconstrained, while most other evaluations have classified skills more granularly (e.g., phonemic awareness as constrained and reading comprehension as unconstrained). In sum, we lack a clear consensus about how to classify skills and measures thereof, making it challenging to compare fade-out trajectories across studies.

Finally, to better understand whether the theory can explain intervention impacts across different types of educational interventions, the theory should be tested in a broad set of experimental data using a priori predictions. The Constrained Skills View has only recently been applied a priori to predict differential patterns of skill development (see McCormick et al., 2017). More commonly, it has been used to advance theoretical claims about intervention effectiveness (Snow & Matthews, 2016) or to make post-hoc explanations about heterogeneous patterns of intervention effects (e.g., Barnett et al., 2018; Durkin et al., 2022).

### *Current Study*

The current study examined whether exogenously produced impacts on unconstrained skills show greater persistence than those on constrained skills. We applied the Constrained Skills View in a meta-analytic dataset of educational intervention RCTs targeting a variety of skills and ages to examine the broad application of this idea across interventions and skills. We systematically categorized measures from 54 studies as constrained or unconstrained before

examining skill-specific impacts. We then tested whether intervention impacts on unconstrained skills persisted more than impacts on constrained skills in the 6- to 12- months after the intervention ended. This test does not allow us to examine whether impacts on unconstrained skills transfer to broader capacities for reading and mathematics in the long term. Instead, our work examines whether intervention impacts are maintained across a broad set of constrained and unconstrained skill measures in the year immediately following the intervention. The treatments in the current sample included a wide variety of educational interventions that vary in their targeted skills(s) and developmental period (infancy/toddlerhood, early childhood, and middle childhood), thus allowing us to examine whether the Constrained Skill Theory extends to skills acquired broadly across educational interventions.

## Methods

### Data

The data used in the current analysis come from the Meta-Analysis of Educational RCTs with Follow-up (MERF; Hart et al., in press) sample. See Hart et al. (in press) for information about sample creation and study coding beyond the scope of this paper. MERF is a meta-analytic dataset comprised of educational randomized controlled trials (RCTs) sampled from the following eight meta-analyses: Bailey et al. (2020), Burns et al. (2016), Kraft et al. (2018), Li et al. (2020), Protzko (2015, 2017), Suggate (2016), and Taylor et al. (2017). These meta-analyses were initially selected to generate a diverse sample of studies that reflected the broad range of educational interventions common in the field (e.g., early childhood programs, adolescent social-emotional interventions, reading remediation). Several of the meta-analyses explicitly analyzed follow-up impacts and/or were highly cited and influential.

These eight meta-analyses yielded 426 unique papers, 400 of which had accessible PDFs in the English language and were then reviewed for inclusion in the MERF sample. The 400 papers reported treatment impacts for 305 unique studies. The research team began the process of winnowing down the paper sample by reviewing the study design of each paper in the aforementioned meta-analyses. We only included RCTs in the sample (i.e., 196 of the 305 studies). Next, we only included studies if the original study team reported treatment impacts on social-emotional and/or cognitive outcomes, which the authorship teams did for 183 studies. We then conducted an extensive follow-up search to gather all available follow-up assessments for each intervention. We only included studies that reported follow-up treatment impacts for the same participants at least 6 months after the posttest. Of the 183 studies, 94 included viable follow-up assessments. Five studies were removed from the sample due to insufficient information to calculate effect sizes, resulting in 89 studies. Finally, the research team removed four nutrition studies as they were not educational in focus and were qualitatively dissimilar from the other studies in the sample. The final sample included 139 papers covering 85 studies (see Figure 1 for sample selection figure).

### *MERF Sample Coding*

A team of three coders (two masters-level research assistants and one PhD student (second author)) double-coded each paper for intervention details and results, with any discrepancies resolved during team meetings. To test reliability, the team first coded ten papers and reached agreement ranging from 82% to 89%. The team coded information about the intervention itself (e.g., duration, intensity, inputs, level of randomization, etc.), treatment and control group details, participant demographics, and information about the measures collected at the posttest and follow-up assessments. Coders recorded the timing of each assessment, with

follow-up assessments coded if they occurred more than six months after the posttest. The coding team also coded information necessary for determining treatment impacts (standard deviations, effect sizes, *p*-values, etc.).

***Calculation of Effect Sizes and Standard Errors***

The coding team then determined posttest and follow-up effect sizes. If the authors reported a viable effect size in standardized units, the team used this effect size (i.e., the effect size came from a model that allowed for the interpretation of an average treatment effect with no interaction or mediators included). If the authors did not report an effect size, the coding team determined the Glass's Delta effect size formula using descriptive reports of group means and standard deviations:

$$ES = \frac{M_{tx} - M_{cntrl}}{sd_{cntrl}}$$

If the author reported both means and standard deviations and effect sizes in standardized units, then the coding team used various criteria (outlined by Hart et al., in press) to determine whether to take the calculated or author-reported effect size, with the ultimate goal of using the best available estimate of the average treatment effect. The coding team used various methods to derive effect sizes when they were not clearly reported in the paper or descriptive statistics were insufficient. These methods included the use of *f*-statistics, *t*-statistics, and *p*-values. In select cases, the author-reported effect sizes varied across papers from the same study or were split by groups (for instance, male versus female). In these cases, the coding team recorded effect sizes for each group and later averaged them to generate overall impact estimates. Adjustments were

made such that higher effects indicated more desirable outcomes (e.g., stronger reading skills, fewer reading problems).

The coding team also used the author-reported *p*-value and standard error for each effect size. If this information was not precisely reported, the coding team calculated the standard error using the following formula:

$$SE_{ES} = \sqrt{\frac{n_{tx} + n_{cntrl}}{n_{tx} n_{cntrl}} + \frac{ES^2}{2(n_{tx} + n_{cntrl})}}$$

The *p*-values were then estimated by calculating the *t*-statistic (effect size divided by standard error) and determining the associated *p*-value. Degrees of freedom were set to the total sample size minus two.

The standard errors included in the current dataset were also adjusted for clustering if they came from interventions that utilized clustered randomization. In cases where we used the author-reported effect size, we assumed the reported standard error was adjusted for clustering. If the standard error was calculated using descriptive information, it was scaled by a variance inflation factor that assumed 20 clusters and an ICC of 0.10.

Follow-up intervention effects were categorized using the following time bins: 6-12 months after the posttest, 1-2 years after the posttest, and greater than 2 years after the posttest. When an intervention reported multiple impacts within the same follow-up window (for instance, an intervention that reported impacts at 6 months and 9 months after the posttest), the effects within that window were averaged.

***Analytic Sample***

Treatment impact estimates are often subject to biased reporting. Common sources of bias include selective reporting of short and long-term effects, wherein larger posttest effects are more likely to be reported, and researchers may be incentivized to collect and report data for these "promising" outcomes at follow-up assessment waves (Bailey et al., 2020; Bailey & Weiss, 2022). Additionally, research teams may change the measures that they collect across assessment waves for a variety of reasons that could be biased toward finding larger, positive effects. Taken together, it is difficult to anticipate how these selection-related dynamics will bias estimates of fadeout and persistence. To generate estimates of fadeout that are less likely to be affected by these selection forces, the current analysis examined fadeout for outcomes that were measured consistently using the same measure, subscale, construct, and reporter at posttest and follow-up within the same treatment-control group contrast.

Thus, the analytic sample was limited to a series of linked effect sizes (subsequently called "aligned groups") for which intervention impacts were reported for the same cognitive construct, measure, subscale, and reporter within the same study and treatment-control group contrast across assessment waves. (Note that some studies had more than one treatment-control group contrast, herein referred to as "interventions," if they included multiple experimental groups). This approach limits the likelihood that complicated selection dynamics may bias our estimates, particularly with regard to researchers changing measures across follow-up assessments. Importantly, however, this method does not avoid all selection bias concerns, as estimates could still be biased by selective posttest reporting and follow-up reporting based on posttest impact magnitude, or conventional *p*-hacking. We attempt to address these concerns related to publication bias and selection into follow-up in sensitivity analyses.

This approach is also helpful because it increases the likelihood that included outcomes are those that researchers anticipated their intervention would affect. In other words, if a researcher spent time and resources collecting data on the same construct and measure at multiple waves, they likely anticipated some link between the outcome and the intervention. This helps in avoiding the inclusion of outcomes that were not theoretically relevant to the intervention model, as it is usually impossible to know which outcomes researchers anticipated would be most affected by the intervention (i.e., "confirmatory" and "exploratory" outcomes are not typically reported).

We use the term "intervention" to refer to unique treatment-control group contrasts. The initial MERF sample included 85 studies comprising 110 interventions with 726 posttest and 1,247 follow-up effect sizes. After limiting the sample to include only aligned groups capturing skills from the same intervention using the same measure, subscale, construct, and reporter at post-test and at least one follow-up, the sample was comprised of 68 studies and 86 interventions with 460 posttest effect sizes and 580 follow-up impacts. From this sample of aligned groups, 236 captured cognitive outcomes, and 214 captured social-emotional outcomes. The average effect for posttests with aligned groups was slightly smaller (0.20 SD) than the impact for posttests without aligned groups (0.29 SD; Hart et al., in press).

**Constrained/Unconstrained Coding**

For the current study, the sample was further limited to only include interventions reporting treatment impacts on aligned groups capturing cognitive skills. Thus, 54 interventions were included, each contributing at least one cognitive skill aligned group (See Figure 1 for sample selection figure). These interventions included literacy interventions ($n$ = 36), early childhood education programs ($n$ = 10 interventions), socioemotional learning interventions ($n$ =

4), math interventions ($n$ = 2), executive functioning interventions ($n$ = 1), and home visiting ($n$ = 1) interventions; See Table S1.

After limiting the sample, a coding team comprised of one Ph.D. student (first author) and two experts in early education coded each combination of construct, measure, and subscale on which impacts were reported as either constrained or unconstrained (See Table S2 for the full list of coded construct, measure, and subscale combinations). The initial MERF sample contained 236 aligned cognitive groups, within which there were 159 unique construct-measure-subscale combinations (e.g., multiple interventions may have used the Woodcock-Johnson Letter Word Identification subtest to measure reading). These 159 construct-measure-subscale combinations were coded as constrained, unconstrained, or exclude (i.e., unable to be coded). Decisions were then applied to all aligned groups for which the measure was used.

Prior to coding, the Ph.D. student compiled definitions of constrained and unconstrained skills from studies on reading achievement and the recent early childhood intervention literature (the definitions used came from Durkin et al., 2022; McCormick & Mattera, 2022; McCormick et al, 2022; McCormick et al., 2020; Paris, 2005; Snow & Matthews, 2016). These definitions were then fed to ChatGPT to summarize the skill types and create one guiding definition that could be used for coding, which we then further edited. Below is the definition that we ultimately used in coding constrained skills:

> Constrained or directly teachable skills are specific competencies that have a ceiling for performance. These skills tend to be acquired more quickly with instruction and can be more readily assessed. Examples of these skills include letter knowledge, rote counting, and phonological awareness. They are important to master during elementary school to ensure success in schooling. Mastery of constrained skills varies over time, with highly

> variable and unstable data distributions during initial acquisition and mastery. These skills tend to have a finite amount of time and attention required for mastery (Adapted from OpenAI, 2023).

The following definition was used for categorizing unconstrained skills:

> Unconstrained skills refer to competencies that develop gradually over time and can never be fully mastered, with no ceiling of perfect performance. These skills are more complex and difficult to assess and include language skills such as vocabulary and reading comprehension and general knowledge of the world. They are acquired through varied experiences and become increasingly crucial to comprehension as the texts become more complex. Unconstrained skills may be particularly important for predicting long-term outcomes. They tend to be more difficult to influence through classroom instruction than constrained skills (Adapted from OpenAI, 2023).

The team first coded 18 of these combinations to assess agreement. Reliability was quantified conservatively as the number of codes agreed upon by all three coders divided by the number of items coded. Reliability in this first phase of coding was 61%. Following the initial coding and discussion of coding disagreement, the doctoral student added lengthier descriptions of each measure and subscale to aid in the coding process. When possible, descriptions were pulled from the intervention papers. When insufficient information was provided, the doctoral student drew text from outside sources (i.e., another paper using the same measure).

The team then began the formal coding process. All unique combinations of construct, measure, and subscale were coded as either constrained, unconstrained, or unclear. Each coder

also rated their confidence in their code using a one to three scale, with one indicating low confidence and three indicating high confidence. Note that descriptive statistics using these confidence ratings came from the second round of coding (coders supplied confidence codes when coding the initial 18 combinations but low confidence ratings were the result of insufficient information about measures). See Table S1 for a full list of the interventions that contributed codes to the constrained, unconstrained, and excluded categories. See Table S2 for a list of constructs, measures, and subscales and their respective final codes.

Across all aligned groups, the three coders had 71% agreement (calculated as the number of codes agreed upon by all three coders divided by the total number of coded combinations of constructs, measures, and subscales). Any codes that all three coders did not agree on were discussed until the coders reached a consensus. If all three coders did not reach a consensus on a code, that construct, measure, and subscale combination was excluded from the analysis. Other reasons for excluding combinations included: insufficient information on the measure, scores comprised of multiple subscales (some of which were constrained and some of which were unconstrained), and measures being unrelated to the theory. In total, 39 construct, measure, and subscale combinations ($n$ = 52 aligned groups) were excluded. (See Table S3 for the breakdown of excluded combinations). If only two of the three coders agreed on a code for a combination ($n$ = 12 aligned groups; see Table S4), the combination was included in a sensitivity analysis but excluded from the primary analyses.

As we describe in more detail below, the majority of effect sizes (91%) included in our analysis were derived from measures of language and literacy. Thus, unconstrained skills tended to capture constructs such as vocabulary (e.g., Expressive One Word Picture Vocabulary Test) and reading comprehension (e.g., Neale Analysis of Reading Ability). However, non-reading

constructs were also included, such as problem-solving strategies for mathematics (e.g., Research-Based Early Math Assessment) and measures of IQ (e.g., Stanford Binet Intelligence Scale). Constrained skills tended to capture phonological measures of reading (e.g., Pollack Tests) and basic numerical operations for math (e.g., Wechsler Applied Problems).

**Analytic Model**

To model fadeout for constrained and unconstrained skills, we employ two approaches. In the first approach, we use a simple weighted average of effect sizes at each assessment wave. We include a random effect for study and weighted estimates by $(\frac{1}{se^2})$ to account for the precision of estimates. This approach follows standard approaches that examine average impacts at each wave to examine trajectories of fadeout.

In our second approach, we applied a simple regression model that assumed an underlying causal pathway by which treatment impacts measured at posttest influence impacts measured at follow-up. Thus, we regressed follow-up effects on posttest effects to determine the extent to which posttest intervention impacts persisted at follow-up. We used a random effects meta-regression in which aligned groups were nested within studies. To account for the nested nature of the data, we included a study random effect in all models. Effects were weighted by $(\frac{1}{SE^2})$ to account for the precision of estimates. Together, by incorporating the variance-covariance matrix of the full model, these methods give greater weight to effect sizes estimated with greater precision, and account for between-study variability adjusted for by the study random effect (Pustejovsky, 2020). To assess whether fadeout differed for unconstrained and constrained skills, we also included a dummy variable for skill type, as well as an interaction between this dummy and the posttest treatment impact. Analyses were conducted using the *metafor* package in R (Viechtbauer, 2010).

The following model was used:

**Level one: Aligned groups**

$$ES_{fsi} = \beta_{0s} + \beta_{1s}ES_{psi} + \beta_2 Unconstrained_{si} + \beta_3 ES_{psi} * Unconstrained_{si} + \boldsymbol{\varepsilon}_{fsi}$$

**Level two: Study**

$$\beta_{0s} = \gamma_{00} + \varphi_{0s}$$

$$\beta_{1s} = \gamma_{10} + \varphi_{1s}$$

Here, the subscript *f* indicates follow-up, *s* indicates study, *i* indicates aligned groups and *p* indicates posttest. Thus, $ES_{fsi}$ is the estimated follow-up effect size for aligned group *i* from study *s* (e.g., an impact on reading comprehension from "Study A" at the 6- to 12-month follow-up window). $ES_{psi}$ refers to the effect size for posttest *p* from study *s* and aligned group *i* (e.g., an impact on the same reading comprehension measure from "Study A" at posttest). $Unconstrained_{si}$ refers to whether group *i* from study *s* is unconstrained (coded as 1) or constrained (coded as 0). In level 2, $\gamma_{00}$ refers to the grand mean intercept across studies, and $\varphi_{0s}$ is the study-specific random effect. We also included a random effect for the slope term, $\varphi_{1s}$, which allows us to examine whether the conditional persistence rates varied by study.

In this model, the slope for the posttest effect size can be understood as the rate of *conditional* persistence between posttest and follow-up. Thus, $\beta_1$ captures the extent to which posttest impacts predict follow-up impacts for constrained skills. Note that we do not normalize our underlying data in any way, so the posttest effects and follow-up effects are scaled in measure-specific standard deviations. As an example, if the observed $\beta_1$ were 1, this would imply complete conditional persistence (i.e., no fadeout) between posttest and follow-up regardless of the initial impact size (assuming zero effect for $\beta_0$ and $\beta_2$), as a posttest effect of 0.25 SDs would

imply a follow-up effect of 0.25 SDs. Alternatively, if $\beta_1$ were 0.50, this would imply that follow-up effects persist at a rate of 50% between posttest and follow-up (i.e., an effect size of 1 SD at posttest would predict a follow-up effect of 0.50 SDs).

This approach has important advantages over approaches that capture fadeout by simply taking the absolute difference between and posttest and follow-up impact in SD units (see Hart et al., in press for additional conceptual discussion of this modeling approach). This approach allows us to observe the *relative rate of persistence* between posttest impacts and follow-up impacts in which the magnitude of posttest intervention effect is essentially controlled. Given that our approach assumes an underlying causal association between posttest effects and follow-up effects whereby larger posttest effects produce larger follow-up effects, examining rates of persistence conditional on posttest impact is critical. In other words, we set an intervention that produced, say, an initial effect of .50 and another that produced an effect of 0.10 on the same footing by asking to what extent their respective follow-up effects persisted as a proportion of these initial effect sizes. In contrast, an "absolute approach" to modeling fadeout might consider a posttest effect size of 0.10 SDs that shrinks by 0.05 SDs to show less fadeout than a posttest effect size of 0.50 SDs that shrinks by 0.10 SDs, when in fact the former represents 50% fadeout and the latter represents 20% fadeout. By examining the extent to which a posttest impact on a specific skill in a specific intervention predicts follow-up impacts, our approach takes into account the many factors across interventions that could produce differences in posttest impact magnitude (e.g., intervention quality, skill type).

The constant term, $\beta_0$, captures the predicted follow-up effect when posttest effects are zero. This term can be thought of as capturing the extent to which interventions produce impacts on follow-up outcomes that are not predicted by posttest effects on that same outcome. These

effects, which can be likened to "sleeper effects," might arise if interventions produce impacts on unmeasured mediators of follow-up effects not fully captured by posttest impacts (Elango et al., 2015; Pages et al., 2022). Such effects might also reflect impacts on developmental processes that do not emerge until later assessments.

The two terms addressing the unconstrained skill codes, $Unconstrained_{si}$ and $ES_{psi}$ * $Unconstrained_{si}$, capture the extent to which these basic parameters differ for unconstrained skills as compared with constrained skills (the referent). Thus, the coefficient for $\beta_2$ captures whether the portion of follow-up intervention impacts unexplained by posttest impact magnitude on the same skill is different for unconstrained skills. Finally, the coefficient for the interaction term, $\beta_3$, captures the extent to which the conditional persistence rate differs between unconstrained and constrained skills. A positive coefficient on either of these two terms would suggest that impacts on unconstrained skills persist at higher rates than constrained skills. Conversely, negative coefficients would suggest that impacts on unconstrained skills fade more rapidly when compared with constrained skills.

### Transparency and Openness

The data and syntax used in the current study will be publicly posted upon acceptance. This study was not preregistered.

## Results

### Descriptive Analyses

Of the 236 aligned groups (impacts reported using the same measure, subscale, and reporter at posttest and at least one follow-up) that measured posttest impacts on cognitive skills, 223 had matched 6- to 12-month follow-up impact estimates, and 28 had matched 1 to 2-year follow-up impact estimates that we could analyze. Our primary analysis focused on follow-ups

collected 6- to 12 months after the posttest assessment. Due to the small number of follow-up assessments collected beyond that period, analyses reporting 1 to 2 years follow-up impacts are reported in the supplement. Of note, there were also 32 aligned groups with follow-up impacts reported past the 2-year follow-up mark. However, the small sample and wide variability in follow-up assessment timing precluded analysis of these effects.

Of the 223 aligned groups with data at the 6- to 12-month follow-up period, 127 were constrained, 48 were unconstrained, and 48 were excluded. Here, we first describe information about the interventions that contributed aligned groups to our sample. Then, we detail the characteristics of the constrained and unconstrained outcomes that comprise these aligned groups.

Table 1 provides intervention and participant characteristics for treatments contributing constrained and unconstrained aligned groups (for characteristics of interventions excluded from this sample, refer to Table S5). As Table 1 reflects, 48 interventions contributed at least one constrained or unconstrained aligned group, 34 interventions contributed at least one constrained aligned group, and 32 interventions contributed at least one unconstrained aligned group. Eighteen interventions contributed both constrained and unconstrained aligned groups. Coders' confidence was generally high and there was little variation by skill type ($M_{constrained}$ = 2.82; $M_{unconstrained}$ = 2.77; range = 2 to 3). The average publication year also varied little by skill type ($M_{constrained}$ = 2007; $M_{unconstrained}$ = 2005), though the range was larger for unconstrained skills. Interventions contributing unconstrained skills had larger ($M_{unconstrained}$ = 445 participants; $M_{constrained}$ = 254 participants) and younger samples ($M_{unconstrained}$ = 65 months; $M_{constrained}$ = 79 months) than did interventions contributing constrained skills. This may indicate that early childhood interventions are more likely to measure unconstrained skills consistently, using the

same measure, at posttest and follow-up. On average, interventions contributing unconstrained skills were longer ($M_{unconstrained}$ = 9.05 months; $M_{constrained}$ = 6.66 months), more intensive ($M_{unconstrained}$ = 133 hours; $M_{constrained}$ = 49 hours), and included more time in school (36% of interventions contributing unconstrained skills and 18% of interventions contributing constrained skills) than interventions contributing constrained skills. It should be noted, however, that as shown in the "N" column of Table 1, many interventions did not report some of this descriptive information.

Most of the studies in the analytic sample targeted reading or language skills, which is reflected in the "treatment focus" panel of Table 1 (84% of interventions contributing unconstrained skills and 100% of interventions contributing constrained outcomes had a reading and language focus). Note that studies could have more than one treatment focus. Math, science, general cognitive skills, learning skills, and social-emotional skills were also targeted by interventions included in our analysis, though to a lesser extent.

[Table 1]

Table 2 details information regarding the outcomes included in the analytic sample (refer to supplemental Table S6 for sample characteristics of the excluded aligned groups). Both constrained (94%) and unconstrained (85%) skills were largely language and literacy outcomes (See column 2). Otherwise, outcomes were taken from measures of math skills, general cognitive skills, and more broad measures of academic functioning. Of note, given that reading and language outcomes comprised a larger portion of the sample, we performed sensitivity analyses in which we limited our analyses to only these outcomes (see more details below). On average, constrained skills had 1.15 follow-up assessments, slightly fewer than the average number of unconstrained skills: 1.49 follow-up assessments.

As shown in columns 4 and 5 of Table 2, we observed larger posttest impacts for constrained skills (0.44 SD) when compared with unconstrained skills (0.34 SD). This pattern was consistent regardless of whether effects were calculated with analytic weights to increase the contribution of more precisely estimated effects. Importantly, Table 2 reflects that these posttest averages were largely driven by the language and literacy outcomes in our sample.

[Table 2]

We also examined the ratio of posttest to follow-up standard deviations by skill type to determine whether the distribution of scores widened or narrowed over time. The Constrained Skills View posits that the variation in constrained skills decreases over time as most children learn these rudimentary skills, though the time frame in which this occurs is unclear (Paris, 2005). In this analysis, we included only raw standard deviations for which scoring did not change between papers in the same study (from standardized to raw and vice versa) and for which the coding team did not need to average across papers or split samples. Our findings (Table S8) suggest that variation in scores widened for both skill types, with the ratio of posttest to follow-up being approximately 1.0 to 1.20 standard deviations for both skill types. The time frame in which variation in constrained skills is expected to winnow is unclear, yet in the short time frame assessed in this study, we do not find evidence of narrowing standard deviations for constrained skills.

**Weighted Average Effect Sizes at Assessment Waves**

In our first modeling approach, we examined average effect sizes at each assessment wave, following typical meta-analytic procedures. Table 3 presents the average weighted effect size across assessment periods through two years follow-up (see supplemental Table S7 for further assessment periods). Across both types of skills, we observed an average posttest effect

of 0.43 SDs that dropped to approximately 0.24 SDs at 6-12 months. For constrained impacts, posttest impacts were 0.44 SDs on average and dropped to 0.27 SDs by the first follow-up wave, whereas the posttest impacts for unconstrained skills dropped from 0.34 SDs to 0.17 SDs. These differences are depicted in Figures 2 and 3 which show average trajectories for constrained skills and unconstrained skills, respectively. Here, the coordinates for each effect size estimate are weighted by the study sample size, reflecting the overall pattern that less precisely estimated treatment impacts were larger than more precisely estimated impacts. The purple line represents the weighted average effect size.

[Table 3]

[Figure 2]

[Figure 3]

**Regression Results**

In our second modeling approach, we regress follow-up effect sizes on post-test to determine the proportion of the post-test impact persisting into follow-up. Table 4 displays the results of the meta-analytic random effects regression model predicting treatment impacts at the 6- to 12-month follow-up assessment by treatment impacts at posttest (treatment impacts at the 1- to 2-year follow-up are shown in Table S9). On average, we observed a conditional persistence rate of 43% for all outcomes at 6- to 12-month follow-up (see column 1), suggesting that follow-up effects were strongly predicted by posttest effects. Without taking the intercept into account, the observed slope term of 0.43 indicates that follow-up impacts were about 43% the magnitude of posttest impacts. The intercept term in this model was small and positive (0.07), providing some evidence that interventions may have effects at follow-up that are not captured by initial posttest impacts, which might be evidence for unmeasured mediators that

produce follow-up impacts unexplained by posttest impacts. However, this term was statistically non-significant, suggesting the estimate is imprecise.

Column 3 presents results from the key model in our study, which included the interaction between the posttest effect and the indicator for whether a skill was constrained or unconstrained. Here, we observed no differences in the small intercept effect by skill type. However, contrary to our predictions, we observed a negative and statistically significant interaction term, $\beta$ = -0.23 ($p$ = 0.01). This suggests that, on average, unconstrained outcomes show a conditional persistence rate that is 23 percentage points *lower* than that of constrained outcomes at 6- to 12-month follow-ups. Figure 4 provides a graphical representation of these findings.

[Figure 4]

***Heterogeneity in Conditional Persistence***

We observed substantial heterogeneity in 6- to 12-month follow-up effect sizes ($I^2$= 61.70%; $\tau_{intercept}$ = 0.24). The introduction of the posttest (Table 4, Baseline) substantially reduced heterogeneity in effect sizes at 6- to 12 months ($I^2$= 4.29%; $\tau_{intercept}$ = 0.11). Neither the introduction of skill type ($I^2$= 1.12%; $\tau_{intercept}$ = 0.11) nor the interaction between skill type and posttest ($I^2$= 0.26%; $\tau_{intercept}$ = 0.11) further contributed to this reduction. The introduction of the random slope term into the model indicated a moderate level of heterogeneity in the model for conditional persistence rates, which was further reduced by the inclusion of the interaction term ($\tau_{slope}$= 0.09 to 0.02).

[Table 4]

**Robustness Checks**

***Inclusion of Relevant Covariates***

We tested the robustness of our findings by adding covariates (Table 5). The correlations between covariates included in the robustness checks are shown in Table S10. These models are intended to determine whether differences in persistence rates are the result of intervention features and confidence in skill type codes as opposed to differences in constrained and unconstrained skills. In all models, we included both the covariate itself and an interaction between the covariate and posttest effect size. All covariates were mean-centered and standardized to ease interpretations of main effects with interactions included. We first included the average confidence ratings for each aligned group (Table 5, Column 1). The introduction of this covariate did not change the results, consistent with the pattern that coders were confident about their codes for both skill types.

Next, we controlled for the intended intervention length given that interventions that measured constrained outcomes were shorter than interventions that measured unconstrained outcomes ($M_{constrained}$ = 6.66 months; $M_{unconstrained}$ = 9.05 months, respectively). For four studies, information on treatment length was not clearly presented in the paper. Thus, for affected outcome groupings (9 unconstrained, 5 constrained), we set missing values to the mean of the samples comprising each skill type (separate means for constrained versus unconstrained skills with data at the 6- to 12- month follow-up). The resulting model was relatively unchanged (see Table 5, Column 2).

We then fit a model controlling for participants' age (in years) pre-intervention (Table 5, column 3). Participants in treatments contributing unconstrained skills were younger than participants in treatments contributing constrained skills ($M_{unconstrained}$ = 5.65 years; $M_{constrained}$ = 6.67 years, respectively). The inclusion of participant age in the model attenuated the difference in conditional persistence rates by skill type ($\beta$ = -0.08, $p$ = 0.57), suggesting that when

controlling for age, the difference in conditional persistence rate by skill type was diminished though less precisely estimated.

Finally, we fit a model in which we controlled for baseline sample size (Table 5, column 4). Treatments contributing unconstrained skills had a larger sample size, on average, compared with treatments contributing constrained skills ($M_{unconstrained}$ = 445.16; $M_{constrained}$ = 254.20, respectively). The introduction of baseline sample size did not change the substantive takeaways from the primary results.

[Table 5]

### *Alternate Modeling Approaches*

To determine whether our results were consistent across alternate reasonable modeling approaches, we ran a series of additional robustness checks shown in the supplemental file. First, we tested a model with an econometric fixed effect for study to compare whether the key model parameters were consistent when looking within studies that contributed both constrained and unconstrained aligned groups (Supplemental Table S11). This model produced findings in the same direction as the primary model, with unconstrained skills showing less conditional persistence than constrained skills. However, the estimated difference in conditional persistence rates was not consistently statistically significant, likely due to limited power on account of the few interventions that contributed aligned groups for both skill types ($n$ = 18 interventions).

We then utilized the Correlated and Hierarchical Model (Pustejovsky & Tipton, 2022) to examine whether our results were consistent when assuming a different dependency structure within our data. This model allows for both between- and within-study variation in effect sizes and assumes there is a single known correlation between effect sizes from the same study. For

this model, we utilized a nested random effects model wherein analytic groups were nested within treatment control group contrasts, which were nested within studies. Our estimates using this model are similar to those of our main model (See Table S12).

Next, we ran our analyses in a series of more restricted and expanded analytic samples (Supplemental Tables S13 to S16). The results of these models were similar to our primary analysis, with unconstrained skills showing more fadeout (though the interaction term was not consistently statistically significant). Table S13 shows the model when only measures with an average confidence rating of 3 were included (indicating all three raters were highly confident in the code). Table S14 shows the model when 12 combinations of construct, measure, and subscale were included for which only two of the three coders agreed on a code (these outcomes were excluded from the primary analyses). Table S15 shows the results for the model when only literacy outcomes were included, as they represented the majority of outcomes in our sample. Table S16 shows a model limited to interventions in which participants were ages seven years and below, given past application of the Constrained Skills View in early childhood intervention research.

**Treatment and Control Group Growth as Proxies for Skill Type**

There is evidence that fadeout of intervention effects is caused by control-group catch-up where students in the counterfactual eventually learn the skills targeted by an intervention after the intervention ends (Bailey et al., 2020; Watts et al., 2024). To test whether this theory could partially explain the elevated levels of persistence for constrained skills, we conducted an additional exploratory analysis. For this analysis, we calculated post-intervention growth levels for both the treatment and control groups. We generated the growth estimates by subtracting posttest means from follow-up means, divided by the control group's posttest standard deviation

(this ensures a consistent scaling for both the treatment and control groups in case some treatments affect the level of variance in the outcome). We estimated growth for aligned groups if (1) means and standard deviations were reported in raw (not standardized) units for the measure, (2) measures did not change units between posttest and follow-up, (3) the MERF coding team did not determine raw means by averaging across groups or papers (see methods), and (4) neither control nor treatment group growth was negative nor greater than 4 standard deviations (unreasonably large).

We then regressed follow-up effects on these growth estimates and post-test effect sizes (in separate models; see Table S17). We also interacted these growth measures with posttest impacts to examine whether conditional persistence rates vary by the magnitude of treatment or control group growth. As Table S17 reflects, estimates were generally small but in the expected direction, with greater control-group growth predicting smaller follow-up effects and greater treatment-group growth predicting larger follow-up effects. However, the estimates were imprecise, and we found no indication that they fully explained the results observed in Table 4.

**Publication Bias**

Various tests of publication bias suggested minimal bias at the post-test and the 6- to 12-month follow-up periods, which are the primary focus of this study, but increased bias at further follow-up assessments. First, we compared the posttest impact for all the coded aligned groups to the posttest effects for aligned groups that had follow-up data. This comparison helps to gauge the extent to which researchers only collected follow-up for measures that had large posttest impacts. Table S18 displays the average estimated posttest effect for aligned groups that are non-missing at each follow-up wave. For constrained and unconstrained skills, we observed little evidence of bias at the 6- to 12-month follow-up assessment. However, we observe evidence of

biased reporting in the unconstrained skills reported more than one year after the interventions (e.g., the average posttest effect for aligned groups at the 1- to 2-years and 2-years follow-up was 0.50 SD compared with the overall posttest effect of 0.34 SD). This suggests evidence of publication bias in that larger effect sizes tend to be followed longer.

Next, we conducted a PEESE Test to examine whether larger standard errors were predictive of larger follow-up effect sizes. First, we regressed post-test effect sizes on post-test standard errors and found that larger standard errors were associated with larger effects ($\beta$ = 2.63, $p$ = 0.01). The same was true for 6-to-12-month follow-up effects when regressed on 6-to-12-month standard errors ($\beta$ = 1.33, $p$ = 0.004). We then included 6- to 12-month standard errors as a covariate in our primary regression model, which suggested a similar pattern (see Table S19). These findings may indicate publication bias in the sample or may suggest that smaller, more intensive interventions are more likely to produce larger impacts.

We next used a series of graphical tests to detect publication bias in our sample. First, we produced funnel plots (Figure S1). Should there be no bias in our sample, the plotted points should fit within the pyramid, with smaller samples producing a smaller effect (Begg & Berlin, 1988). At posttest, there was minimal evidence of publication bias (both smaller and larger studies report small and large effects). However, at the 6-to-12-month follow-up and onward, the data is skewed to the right, suggesting small studies with small effects were likely under-reported. We then created these same plots separately for studies contributing constrained and unconstrained skills (Figures S2 – S3). When split along these dimensions, evidence of bias appeared similar across interventions contributing each skill type.

We also plotted the percentage of $p$-values less than 0.05 at each assessment point (see Figure S4). We were particularly interested to observe the proportion of values near 0.05, which

could be indicative of *p*-hacking. At the posttest, there was little evidence of *p*-hacking, as there were few *p*-values just below 0.05 (Simonsohn et al., 2015). At the 6- to 12-month and 1 to 2-year follow-ups, however, a greater proportion of *p*-values were just below 0.05, suggesting evidence of selective reporting. At greater than 2 years follow-up, only three of 28 effect sizes had associated *p*-values below 0.05, indicating a high rate of failure to reject the null of zero intervention impact. We then created these plots separately for aligned groups coded as constrained and unconstrained (Figures S5 – S6). Among aligned groups coded as constrained, we saw potential evidence of selective reporting at the greater than 2-year follow-up, though it should be noted only 15 aligned groups reported effect sizes at this timepoint. Among aligned groups coded as unconstrained, we found the strongest evidence of selective reporting at the 1 to 2-year follow-up assessment.

## Discussion

Researchers have argued that skill development can be better understood by considering distinctions between constrained skills, which have a ceiling and a finite endpoint for mastery, and unconstrained skills, which have no clear ceiling and continue to develop over extended periods (Paris, 2005; Snow & Matthews, 2016). These classifications have been applied to literacy (Grøver et al., 2024; Suggate, 2016) and early childhood intervention studies (McCormick et al., 2021; McCormick & Mattera, 2022) to predict the maintenance or fadeout of follow-up impacts, with some arguing that impacts on unconstrained skills should persist more than impacts on constrained skills because they are less likely to develop in counterfactual conditions. However, to date, there has not been a systematic cross-study evaluation of these dynamics in the immediate post-program year and in a broader set of interventions.

In the current study, we expanded the Constrained Skills View by examining whether intervention impacts on constrained skills showed more fadeout in the short term than impacts on

unconstrained skills across a range of educational intervention evaluations. In a meta-analytic dataset of educational RCTs, we systematically classified combinations of constructs, measures, and subscales as either constrained, unconstrained, or excluded from analyses (i.e., not codable for this construct). First, we determined average impacts from the posttest to 6–12-month follow-up assessment. We then operationalized conditional persistence as the proportion of the posttest impact that persisted at the follow-up assessment using a regression-based approach. This method allowed us to estimate the relative persistence of posttest impacts of different magnitudes.

Overall, we found no evidence to suggest that impacts on unconstrained skills persisted at a higher rate than impacts on constrained skills in the year following intervention end. Instead, we found some evidence suggesting the opposite: constrained skills produced an average conditional persistence rate of 48%, whereas unconstrained skills produced an average persistence rate of 25%. The direction of these effects was similar across most sensitivity checks, though the difference in conditional persistence rates was not consistently statistically significant. Below, we consider what these results suggest for theory regarding skill trajectories and intervention design, and we also consider limitations and future directions.

Several findings from the current study are useful for the field as scholars continue to explore the relevance of Constrained Skill Theory for forming predictions about the combinations of skills and interventions that will produce persistent impacts. First, the results of our coding process suggest challenges in determining whether skills are constrained or unconstrained. Before discussion, the coders had a reliability of 71% which, although acceptable, still indicates the three coders did not initially agree on codes for nearly a third of the measures (when comparing reliability between only two coders, reliability ranged from 77% to 81%).

Although this level of reliability might be explained by features of the coding process, it may also indicate challenges in determining whether measures are constrained or unconstrained. Complexity in coding was partly due to the fact that some measures tapped both constrained and unconstrained skills. For example, one of the measures with coder disagreement was the *Overall Quotient* score of the Gray Oral Reading Test. This score includes skills that might be considered constrained (reading speed) and skills that might also be considered unconstrained (reading comprehension). Although such composite scores were generally excluded from analysis, this example highlights that while some skills might be clearly constrained or unconstrained in theory, measures of achievement often include elements of both, rendering it difficult to apply the Constrained Skills View consistently.

The definition of constrained skills as skills with a clear ceiling further complicates operationalization. In many ways, Paris' (2005) original work laying out the constrained versus unconstrained distinction was a methodological critique of the reading achievement literature. Paris was largely concerned with the *measurement* of constrained skills, as measures of finite skills would have insufficient variance for linear statistical analyses once children reach mastery and hit the ceiling on the measure. Although we coded 131 skill/measure combinations as "constrained" in concept, we had little empirical evidence that any of the measures included in the data were actually constrained in the ways that Paris described, an observation made by Suggate (2016) as well. Indeed, we found that the ratio of follow-up SDs to posttest SDs was approximately 1.20 across both skill types (See Table S8), suggesting widening variation for both constrained and unconstrained skills over time. Of course, it should be noted that almost all of our analyses were limited to follow-up periods within a year of the initial post-test. Paris (2005) did not specify the time frame in which winnowing of the distribution of constrained

skills would occur, as it would be expected to differ across specific skills. It is also possible that many of these constrained measures would eventually hit ceiling points that are outside the developmental range included in the studies in our sample. Yet, this relatively short follow-up period (i.e., 1 year) is the time frame in which most intervention studies observe substantial fadeout of effects (e.g., Bailey et al., 2017), and it is similar to the periods used in other applications of this theory in the literacy (e.g., Grøver et al., 2024; Suggate, 2016) and early childhood intervention (e.g., McCormick et al., 2022) literature.

These considerations bear little on the theoretical plausibility of the Constrained Skills View as a partial account for why fadeout happens; however, they help clarify the difficulty associated with using the Constrained Skills View to make clear predictions about whether constrained or unconstrained skills should show higher persistence. When considering the measurement issues in the application of the Constrained Skills View for empirical work, the Woodcock-Johnson Letter-Word ID subtest is an apt example of the difficulties one encounters. The measure is often considered constrained because it taps letter naming at early ages (e.g., Johnson et al., 2022). However, the measure itself has no clear ceiling; it was normed to be administered through late adulthood. The easiest items entail recognition of letters and simple words including “me” and “red,” while the later items on the test entail reading words like “septuagenarian” and “coiffure” aloud. However, the measure contains few items at each developmental stage, and it seems reasonable to expect that the instructional context for a given child could impose a kind of ceiling on the child’s growth at any given time point. If a child received no reading instruction in kindergarten that allowed them to move beyond reading monosyllabic words, then we might still expect to observe the “constrained skill effect” on the Letter-Word ID subtest despite the lack of a true ceiling on the measure. Some evidence from

other constrained measures, however, suggests this is not the case. ECE applications of this theory using the DIBELS Letter Naming Fluency subtest (often considered a constrained measure that is normed through eighth grade) have still reported widening variances during the one-year follow-up period (e.g., McCormick et al., 2021), suggesting substantial growth on the measure with little evidence for ceiling effects. A recent examination of fadeout following Pre-K in North Carolina on composite scores on the DIBELS measure of literacy also found a similar pattern of growth following the intervention (Carr et al., 2024).

The current study also raises questions about the perceived "fundamentality" of constrained and unconstrained skills and implications for fadeout dynamics. Following the conceptualization of "trifecta skills" laid out by Bailey et al. (2017), several studies have argued that unconstrained skills might meet the criteria for trifecta skills due to their fundamentality. As Bailey et al. argued, skills that are more fundamental to further skill development should show greater persistence following an intervention. However, our findings complicate this possibility, as the unconstrained skill impacts faded more quickly despite the apparent fundamental nature of most of the skills tapped by the unconstrained measures (see Supplement Table S2 for a full list). Contrary to expectations, the fundamentality of the skill could also undermine the persistence rate if fundamental skills receive more instruction after the intervention ends in counterfactual conditions, generating increased control group catch-up. For example, reading comprehension, an unconstrained skill, is certainly targeted in children's typical educational environments. However, many of the reading skills captured by the constrained category, like phoneme segmentation, might not receive direct instruction in the counterfactual. Thus, students who receive a reading intervention that teaches phoneme segmentation may continue to show persistent effects on this specific skill. Conversely, students who receive a reading

comprehension intervention may not outperform their control-group counterparts in the long run as the control group has continued opportunities to practice comprehension.

Should this theory be true, control group catchup may explain the unexpected results of the current study. While the Constrained Skills View implies that schools tend to focus on constrained skills (Snow & Matthews, 2016), it is possible that some unconstrained skills may also receive substantial instructional attention. To our knowledge, only one study has examined this in kindergarten and found that math instruction tended to be more constrained and literacy content tended to be slightly more unconstrained, though in a district in which significant curriculum reforms aimed at increasing unconstrained skill instruction (McCormick et al., 2022). Other work has examined the amount of time PreK classrooms in another district spend in "meaning-focused" (e.g., instruction in expressive and receptive vocabulary and listening comprehension) versus "code-focused instruction" (e.g., instruction in letters, rhyming, and letter-sound correspondence) and found classrooms spend more time in meaning-focused literacy activities (Connor et al., 2006). Should unconstrained skills receive instructional attention, the control group's school experience may provide opportunities for unconstrained skill growth, thus attenuating the treatment effect over time. We tested this theory by examining whether the conditional persistence rate varied by levels of control and treatment group growth (See Table S17). Although effects were in the expected direction, with greater control group growth predicting smaller follow-up effects and greater treatment growth predicting larger follow-up effects, estimates were imprecisely estimated making it difficult to draw strong conclusions. Another potential explanation for our surprising finding is that, because unconstrained skills consist of more components (by definition), test content for the same measure may vary more across years as students progress when compared with constrained tasks. If true, then impacts on

constrained skills may more strongly affect performance on subsequent tests, even in the absence of transfer in the post-treatment period. This explanation would also not rule out the Constrained Skills View as a potential explanation for fadeout. However, it also highlights the difficulty with mapping the Constrained Skills View onto different rates of predicted persistence across measures.

**Limitations**

The current study had multiple limitations that deserve attention. First, our sample consisted predominantly of reading and literacy measures, limiting our ability to examine the Constrained Skills View as it applied to other cognitive domains. Although our results were consistent when we limited our sample to language and literacy measures, it is unclear whether these impacts would replicate in a sample consisting entirely of other academic abilities, like math or science.

The second limitation of the current study is the small number of aligned groups with data past the 6- to 12-month follow-up. Only 19 aligned groups had data at the 1 to 2-year follow-up assessment and were coded as either constrained or unconstrained (the rest were excluded). Although we report results from these aligned groups in the supplement (Table S9), our confidence in these estimates is hindered by the small sample size. The short timeframe included in the current study prevents us from examining whether skill trajectories may change in long-term follow-ups. Importantly, Paris (2005) did not specify the time frame in which unconstrained skills should be expected to show widening variation and greater influence on long-term reading abilities. Perhaps in the long-term, the control group masters constrained skills while showing a slower rate of learning on unconstrained skills as time progresses. We should note, however, that the small number of assessments at further follow-up waves represents the

state of the education literature (Watts et al., 2019) and the typical follow-up period reported in studies applying the Constrained Skills View (Grøver et al., 2024; McCormick et al., 2021, 2022; McCormick & Mattera, 2022; Suggate, 2016). Furthermore, evidence suggests most fadeout occurs in the months following an intervention (Bailey et al., 2017; Bailey et al., 2020). Although this limitation inhibits our ability to make predictions about long-term trajectories of constrained and unconstrained skills, we observed evidence of a similar pattern at the 1- to 2-year follow-up with the limited data available.

The third limitation of the current study is our inability to make statements about the validity of the constrained skill view in specific interventions and age groups. Our study did not code the content of interventions along the constrained/unconstrained continuum, nor do we know what instruction was provided to the control group in each study. The alignment between intervention content and skill focus may also influence the maintenance of treatment impacts, but that was beyond the scope of the current study (the existing evidence that skill and content alignment predicts persistence in the literacy meta-analyses discussed in the introduction is mixed). Thus, it remains possible that there are certain age groups, intervention types, outcome measures, and counterfactual conditions for which the interactions of these factors will produce results that match the predictions of the Constrained Skills View. However, meta-analyses are designed to assess broad patterns of effects in a literature, making such interaction effects beyond the scope of what we can observe. We would encourage further causally-informed research that can capture more detailed combinations of intervention foci and outcomes in the long term. However, we would stress the importance of making firm a priori predictions when evaluating complex interactions across intervention features and outcomes.

The final limitation of the current study is our inability to make inferences about skill transfer. Despite our finding that unconstrained skills show greater fadeout, we caution against concluding that unconstrained skills do not support the development of other skills. Paris (2005) suggests that unconstrained literacy skills might be more important for predicting broad reading ability than constrained skills. In the current study, we examined fadeout among skills assessed with the same construct, measure, and subscale over time. We cannot determine whether intervention impacts on constrained or unconstrained skills are likely to transfer to support functioning in other domains. Our findings do not rule out the possibility that impacts on unconstrained skills could be more important for the development of reading ability than impacts on constrained skills, despite evidence of fadeout in both domains. Rather, our findings suggest that both constrained and unconstrained skills measured the same way over time show fadeout, and unconstrained skills show greater fadeout. It remains possible that both skill types are pivotal to the development of more broad skills, such as comprehension or general cognitive ability.

**Conclusion**

In the current study, we extended the Constrained Skills View to a wide variety of interventions and developmental periods to examine short-term trajectories of fadeout. Our findings suggest evidence of fadeout for both constrained and unconstrained skills. Such results align with results from another paper using the MERF dataset that reported fadeout for both social-emotional and cognitive impacts (Hart et al., in press). These findings do not exclude the possibility of transfer but suggest short-term fadeout on the same skills measured across time, irrespective of skill type.

References

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*, 1103–1127. https://doi.org/10.1037/bul0000212

Ackerman, P. L. (2007). New Developments in Understanding Skilled Performance. *Current Directions in Psychological Science*, *16*(5), 235–239. https://doi.org/10.1111/j.1467-8721.2007.00511.x

Ansari, A., Zimmermann, K., Pianta, R. C., Whittaker, J. V., Vitiello, V. E., Yang, Q., & Ruzek, E. A. (2023). The First-Grade Outcomes of Pre-K Attendees: Examining Benefits as a Function of Skill Type, Environments, and Subgroups. *American Educational Research Journal*, *60*(6), 1139–1173. https://doi.org/10.3102/00028312231195559

Bailey, D., 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., 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. & Weiss, M. J. (2022). *Do meta-analyses oversell the longer-term effects of programs? (Part 1). Detecting follow-up selection bias in studies of postsecondary education programs* [MDRC blog post]. MDRC.

Barnett, W. S., Jung, K., Friedman-Krauss, A., Frede, E. C., Nores, M., Hustedt, J. T., Howes, C., & Daniel-Echols, M. (2018). State Prekindergarten Effects on Early Learning at

Kindergarten Entry: An Analysis of Eight State Programs. *AERA Open*, *4*(2), 2332858418766291. https://doi.org/10.1177/2332858418766291

Begg, C. B., & Berlin, J. A. (1988). Publication Bias: A Problem in Interpreting Medical Data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *151*(3), 419–463. https://doi.org/10.2307/2982993

Burns, M. K., Petersen-Brown, S., Haegele, K., Rodriguez, M., Schmitt, B., Cooper, M., Clayton, K., Hutcheson, S., Conner, C., Hosp, J., & VanDerHeyden, A. M. (2016). Meta-analysis of academic interventions derived from neuropsychological data. *School Psychology Quarterly*, *31*(1), 28–42. https://doi.org/10.1037/spq0000117

Carpenter, T. P., Franke, M. L., Jacobs, V. R., Fennema, E., & Empson, S. B. (1998). A Longitudinal Study of Invention and Understanding in Children's Multidigit Addition and Subtraction. *Journal for Research in Mathematics Education*, *29*(1), 3–20. https://doi.org/10.2307/749715

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*, *95*(4), 1200–1217. https://doi.org/10.1111/cdev.14076

Claessens, A., Engel, M., & Curran, F. C. (2014). Academic Content, Student Learning, and the Persistence of Preschool Effects. *American Educational Research Journal*, *51*(2), 403–434. https://doi.org/10.3102/0002831213513634

Connor, C. M., Morrison, F. J., & Slominski, L. (2006). Preschool instruction and children's emergent literacy growth. *Journal of Educational Psychology*, *98*(4), 665. https://doi.org/10.1037/0022-0663.98.4.665

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*, 470–484. https://doi.org/10.1037/dev0001301

Elango, S., García, J. L., Heckman, J. J., & Hojman, A. (2015). Early childhood education. *National Bureau of Economic Research*. http://doi.org/10.3386/w21766.

Fikrat-Wevers, S., van Steensel, R., & Arends, L. (2021). Effects of family literacy programs on the emergent literacy skills of children from low-SES families: A meta-analysis. *Review of Educational Research*, *91*(4), 577-613. https://doi.org/10.3102/0034654321998075.

Grøver, V., Gustafsson, J.-E., Rydland, V., & Snow, C. E. (2024). Are There Sustained Effects of a Preschool Shared-Reading Intervention Addressing Dual Language Learners? *Scientific Studies of Reading*, *28*(4), 441–462. https://doi.org/10.1080/10888438.2024.2335925

Gu, Q., Zou, L., Loprinzi, P. D., Quan, M., & Huang, T. (2019). Effects of open versus closed skill exercise on cognitive function: a systematic review. *Frontiers in Psychology*, 10, 467457. https://doi.org/10.3389/fpsyg.2019.01707

Hart, E. R., Bailey, D. H., Luo, S., Sengupta, P., & Watts, T. W. (in press). 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.*

Johnson, A.D., Partika, A., Martin, A., Lyons, I., Castle, S., Phillips, D.A., & The Tulsa SEED Study Team. (2022). Following the preschool boost into third grade: Do public preschool benefits on cognitive and self-regulatory skills persist? (Child Development & Social Policy Lab Working Paper #2). Georgetown University.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. *Review of Educational Research*, *88*(4), 547–588. https://doi.org/10.3102/0034654318759268

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 (Ed Working Paper No. 20-201). Annenberg Institute for School Reform at Brown University. http://doi.org/10.26300/5tvg-nt21.

Mattera, S., Jacob, R., & Morris, P. (2018). *Strengthening Children's Math Skills with Enhanced Instruction: The Impacts of Making Pre-K Count and High 5s on Kindergarten Outcomes* (SSRN Scholarly Paper 3167484). https://papers.ssrn.com/abstract=3167484

McCormick, M., Hsueh, J., Weiland, C., & Bangser, M. (2017). The Challenge of Sustaining Preschool Impacts: Introducing ExCEL P-3, a Study from the Expanding Children's Early Learning Network. In *MDRC*. MDRC. https://eric.ed.gov/?id=ED575653

McCormick, M., & Mattera, S. (2022). Learning More by Measuring More: Building Better Evidence on Pre-K Programs by Assessing the Full Range of Children's Skills. Measures for Early Success. *MDRC*.

McCormick, M., Pralica, M., Weiland, C., Hsueh, J., Moffett, L., Guerrero-Rosada, P., Weissman, A., Zhang, K., Maier, M. F., Snow, C. E., Davies, E., Taylor, A., & Sachs, J. (2022). Does kindergarten instruction matter for sustaining the prekindergarten (PreK) boost? Evidence from individual- and classroom-level survey and observational data. *Developmental Psychology*, *58*(7), 1298. https://doi.org/10.1037/dev0001358

McCormick, M., Weiland, C., Hsueh, J., Pralica, M., Weissman, A. K., Moffett, L., Snow, C., & 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.

McCormick, M., Weissman, A. K., Weiland, C., Hsueh, J., Sachs, J., & Snow, C. (2020). Time well spent: Home learning activities and gains in children's academic skills in the prekindergarten year. *Developmental Psychology*, *56*(4), 710. https://doi.org/10.1037/dev0000891

OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat

Pages, R., Bailey, D. H., & Duncan, G. J. (2022). *Exploring the "Dark Matter" of Early Childhood Educational Programs: A Pattern-of-Indirect-Effect Approach* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/zubg8.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly*, *40*(2), 184–202. https://doi.org/10.1598/RRQ.40.2.3

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. (2017). Raising IQ among school-aged children: Five meta-analyses and a review of randomized controlled trials. *Developmental Review*, *46*, 81–101. https://doi.org/10.1016/j.dr.2017.05.001.

Pustejovsky, J. E. (2020). *Weighting in multivariate meta-analysis*. https://www.jepusto.com/weighting-in-multivariate-meta-analysis/

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with Robust Variance Estimation: Expanding the Range of Working Models. *Prevention Science*, *23*(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3

Rittle-Johnson, B., Schneider, M., & Star, J. R. (2015). Not a one-way street: Bidirectional relations between procedural and conceptual knowledge of mathematics. *Educational Psychology Review*, *27*, 587-597. https://doi.org/10.1007/s10648-015-9302-x.

Silverman, R. D., Johnson, E., Keane, K., & Khanna, S. (2020). Beyond Decoding: A Meta-Analysis of the Effects of Language Comprehension Interventions on K–5 Students' Language and Literacy Outcomes. *Reading Research Quarterly*, *55*(S1), S207–S233. https://doi.org/10.1002/rrq.346

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

Snow, C. E., & Matthews, T. J. (2016). Reading and Language in the Early Grades. *The Future of Children*, *26*(2), 57–74. https://doi.org/10.1353/foc.2016.0012

Spiegel, J. A., Goodrich, J. M., Morris, B. M., Osborne, C. M., & Lonigan, C. J. (2021). Relations between executive functions and academic outcomes in elementary school children: A meta-analysis. *Psychological Bulletin*, *147*(4), 329–351. https://doi.org/10.1037/bul0000322

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

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting Positive Youth Development Through School-Based Social and Emotional Learning Interventions: A Meta-Analysis of Follow-Up Effects. *Child Development*, *88*(4), 1156–1171. https://doi.org/10.1111/cdev.12864.

Viechtbauer W. (2010). "Conducting meta-analyses in R with the metafor package." *Journal of Statistical Software*, 36(3), 1–48. doi:10.18637/jss.v036.i03.

von Hippel P. T. (2015). The heterogeneity statistic I(2) can be biased in small meta-analyses. *BMC medical research methodology*, *15*, 35. https://doi.org/10.1186/s12874-015-0024-z

Watts, T.W., Botvin, C.M., Bailey, D.H., Hart E.R., Mattera S., Clements, D.H., Sarama, J., Farran, D., & Lipsey, M.W. (2024). Predicting Persistence and Fadeout Across Multi-Site RCTs of an Early Childhood Mathematics Curriculum Intervention. [Manuscript Submitted for Publication].

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

Whittingham, C. E., Hoffman, E. B., & Paciga, K. A. (2021). Assessment, accountability, and access: Constrained skill mastery as instructional gatekeeper. *Journal of Early Childhood Literacy*, 14687984211042056. https://doi.org/10.1177/14687984211042056

Table 1
Study-Level Intervention and Participant Characteristics (mean [minimum- maximum])

| | **Constrained or Unconstrained (1)** | | **Constrained (2)** | | **Unconstrained (3)** | |
|---|---|---|---|---|---|---|
| | **Mean** | ***N*** | **Mean** | ***N*** | **Mean** | ***N*** |
| Base sample size | 330.67 [24 - 3933] | 48 | 254.20 [24.00 - 3933] | 34 | 445.16 [30.00 - 3933] | 32 |
| Confidence rating | 2.77 [2.00 - 3.00] | 47 | 2.82 [2.33 - 3.00] | 33 | 2.77 [2.00 - 3.00] | 32 |
| Publication year | 2005 [1969 - 2022] | 48 | 2007 [1983 - 2014] | 34 | 2005 [1969 - 2022] | 32 |
| Baseline age (months) | 67.88 [0.00 - 122.00] | 48 | 78.53 [42.00 - 122.00] | 34 | 64.66 [0.00 - 122.00] | 32 |
| Intended months of treatment | 8.22 [0.92 - 36.00] | 42 | 6.66 [1.15 - 15.50] | 31 | 9.05 [0.92 - 36.00] | 26 |
| Intended hours of treatment | 90.59 [4.00 - 977.62] | 35 | 48.58 [4.00 - 162.94] | 30 | 132.51 [5.00 - 977.62] | 19 |
| Included extra time in school (%) | 23.81 | 42 | 17.86 | 28 | 35.71 | 28 |
| **Treatment Focus (%)** | | | | | | |
| Math | 8.33 | 48 | 2.94 | 34 | 12.50 | 32 |
| Reading/Language | 89.58 | 48 | 100 | 34 | 84.38 | 32 |
| Science | 2.08 | 48 | 0.00 | 34 | 3.12 | 32 |
| General cognitive | 10.42 | 48 | 2.94 | 34 | 15.62 | 32 |
| Learning skills | 2.08 | 48 | 0.00 | 34 | 3.12 | 32 |
| Social-emotional skills | 20.83 | 48 | 8.82 | 34 | 31.25 | 32 |
| **Adult Involvement (%)** | | | | | | |
| Teacher | 56.25 | 48 | 52.94 | 34 | 62.5 | 32 |
| Parent | 20.83 | 48 | 11.76 | 34 | 28.12 | 32 |
| Female participants (%) | 44.16 | 38 | 42.16 | 26 | 44.27 | 27 |

*Note*. Intervention characteristics for unique treatment-control group contrasts that contributed to each of the coding categories (constrained versus unconstrained) are presented (see Table S5 for intervention and participant characteristics of excluded interventions). Column 1 shows characteristics for interventions contributing aligned groups capturing constrained and/or unconstrained skills (interventions could contribute both constrained and unconstrained skills if they measured a variety of skills). Column 2 shows the characteristics of interventions contributing aligned groups capturing constrained skills. Finally, Column 3 shows the characteristics of interventions contributing aligned groups capturing unconstrained skills. "N" indicates the number of interventions that reported information. Eighteen interventions contributed both unconstrained and unconstrained skills and are thus represented in each of the columns above (hence why the sum of Columns 1 and 2 is not equal to the sum of Column 3). Eighteen combinations of construct, measure, and subscale were initially coded as coders learned the coding scheme. These 18 combinations came from seven interventions (one of which contributed a constrained aligned group and all of which contributed excluded aligned groups). One treatment that contributed a constrained aligned group was therefore excluded from the descriptive statistic of mean confidence rating.

Table 2
Analytic Sample Characteristics for Coded Outcomes

| | Treatment groups (#) (1) | Aligned groups (#) (2) | Avg # follow-ups (3) | Avg Posttest ES, weighted (SE) (4) | Avg Posttest ES, unweighted (SE) (5) |
|---|---|---|---|---|---|
| **Constrained or Unconstrained** | **48** | **184** | **1.24** | **0.43 (0.06) *** ** | **0.42 (0.06) *** ** |
| Language and Literacy | 41 | 168 | 1.20 | 0.45 (0.07) *** | 0.44 (0.07) *** |
| Math | 9 | 9 | 1.44 | 0.14 (0.18) | 0.12 (0.18) |
| Cognitive | 3 | 3 | 2.67 | 0.56 (0.13) | 0.57 (0.09) * |
| Other Academic Ability | 3 | 3 | 1.00 | 0.12 (0.03) *** ° | 0.12 (0.03) *** ° |
| Achievement Composite | 1 | 1 | 2.00 | | 0.21 (0.05) ❑ |
| **Constrained** | **34** | **131** | **1.15** | **0.44 (0.07) *** ** | **0.43 (0.08) *** ** |
| Language and Literacy | 31 | 123 | 1.14 | 0.45 (0.07) *** | 0.45 (0.08) *** |
| Math | 7 | 7 | 1.14 | 0.00 (0.20) | 0.02 (0.22) |
| Achievement Composite | 1 | 1 | 2.00 | | 0.21 (0.05) ❑ |
| **Unconstrained** | **32** | **53** | **1.49** | **0.34 (0.06) *** ** | **0.38 (0.09) ** ** |
| Language and Literacy | 25 | 45 | 1.40 | 0.35 (0.08) *** | 0.38 (0.10) ** |
| Cognitive | 3 | 3 | 2.67 | 0.56 (0.13) | 0.57 (0.09) * |
| Math | 2 | 2 | 2.5 | 0.48 (0.05) *** ° | 0.47 (0.05) *** ° |
| Other Academic Ability | 3 | 3 | 1.00 | 0.12 (0.03) *** ° | 0.12 (0.03) *** ° |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

*Note*: Treatments could have more than one content focus, hence why values in column 1 do not add up to the bolded value. The number of aligned groups refers to the number of groupings that included a posttest and at least one follow-up assessment of the same construct measured using the same measure, subscales, and reporter, within a treatment-control group contrast. Average number of follow-ups reflects the average number of follow-up assessments that were collected (at least 6 months after the posttest). "ES" stands for effect size (in standard deviation units). "Constrained or unconstrained" does not include excluded aligned groups. Effects were estimated in R using the "metafor" package. Weighted effects were estimated using a study-level random effect, weighting, and robust standard errors. Unweighted effects were determined using a fixed-effects meta-analytic model with no random or econometric fixed effect for study ID.

° indicates cluster-robust standard errors could not be produced because all outcomes came from one study

❑ indicates that only one aligned group contributed an outcome and we could not determine a weighted or unweighted average. The estimate shown in column 5 is the effect size and standard error for the one aligned group.

Table 3
Average Weighted Meta-Analytic Effect Sizes at Each Binned Study Wave

| | Constrained or Unconstrained | | Constrained | | Unconstrained | |
|---|---|---|---|---|---|---|
| | **Avg ES (SE)** | ***n*** | **Avg ES (SE)** | ***n*** | **Avg ES (SE)** | ***n*** |
| Posttest | 0.43 (0.06) *** | 184 | 0.44 (0.07) *** | 131 | 0.34 (0.06) *** | 53 |
| 6 months to 1 year | 0.24 (0.04) *** | 175 | 0.27 (0.05) *** | 127 | 0.17 (0.04) ** | 48 |
| >1 year, up to 2 years | 0.19 (0.07) * | 19 | 0.20 (0.10) | 8 | 0.14 (0.08) | 11 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
*Note*: "ES" = effect sizes. Effect sizes are in standard deviation units. The analytic sample was constituted of aligned groups (groupings that included a posttest and at least one follow-up assessment of the same construct measured using the same measure, subscales, and reporter within a treatment-control contrast). Note that "constrained or unconstrained" refers to outcomes coded as either constrained or unconstrained (not excluded). Estimates were estimated in R using the "metafor" package and included a study-level random effect, weighting, and robust standard errors.

Table 4
Modeled Conditional Persistence Rates for Unconstrained & Constrained Outcomes at 6- to 12- Months Follow-Up

| | Baseline (1) | Skill Type (2) | Interaction (3) |
|---|---|---|---|
| Intercept (null = 0.27 (0.04) ***) | 0.07 (0.03) | 0.08 (0.04) | 0.07 (0.03) |
| Posttest | 0.43 (0.05) *** | 0.42 (0.05) *** | 0.48 (0.04) ** |
| Unconstrained | | -0.03 (0.02) | 0.02 (0.02) |
| Unconstrained * Posttest | | | -0.23 (0.07) * |
| Random intercept (null = 0.24) | 0.11 | 0.11 | 0.11 |
| Random slope | 0.09 | 0.10 | 0.02 |
| $I^2$ (null = 61.70%) | 4.29% | 1.12% | 0.26% |
| *N* (Aligned groups/Treatments/Studies) | 175/43/31 | 175/43/31 | 175/43/31 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

*Note*. The unit of analysis is aligned groups of posttest impacts and impacts at 6-to-12 months after the posttest wherein the construct, measure, subscale, and reporter were the same. The "null" model indicates a model estimating the average impact at the 6 – 12-month follow-up. Models were fit using the "metafor" package in R, with a study-level random effect, random slopes for posttest effect size, weighting, and robust standard errors. Standard errors are presented in parentheses, and coefficients can be interpreted in effect size units (i.e., the underlying data are in the effect size units reported by studies). Unconstrained refers to an indicator variable wherein "1" indicates an unconstrained aligned group and "0" indicates a constrained group. Heterogeneity statistics are shown for the null model and each of the subsequent models

Table 5
Modeled Conditional Persistence Rates for Unconstrained & Constrained Outcomes at 6- to 12- Months Follow-Up with Covariates

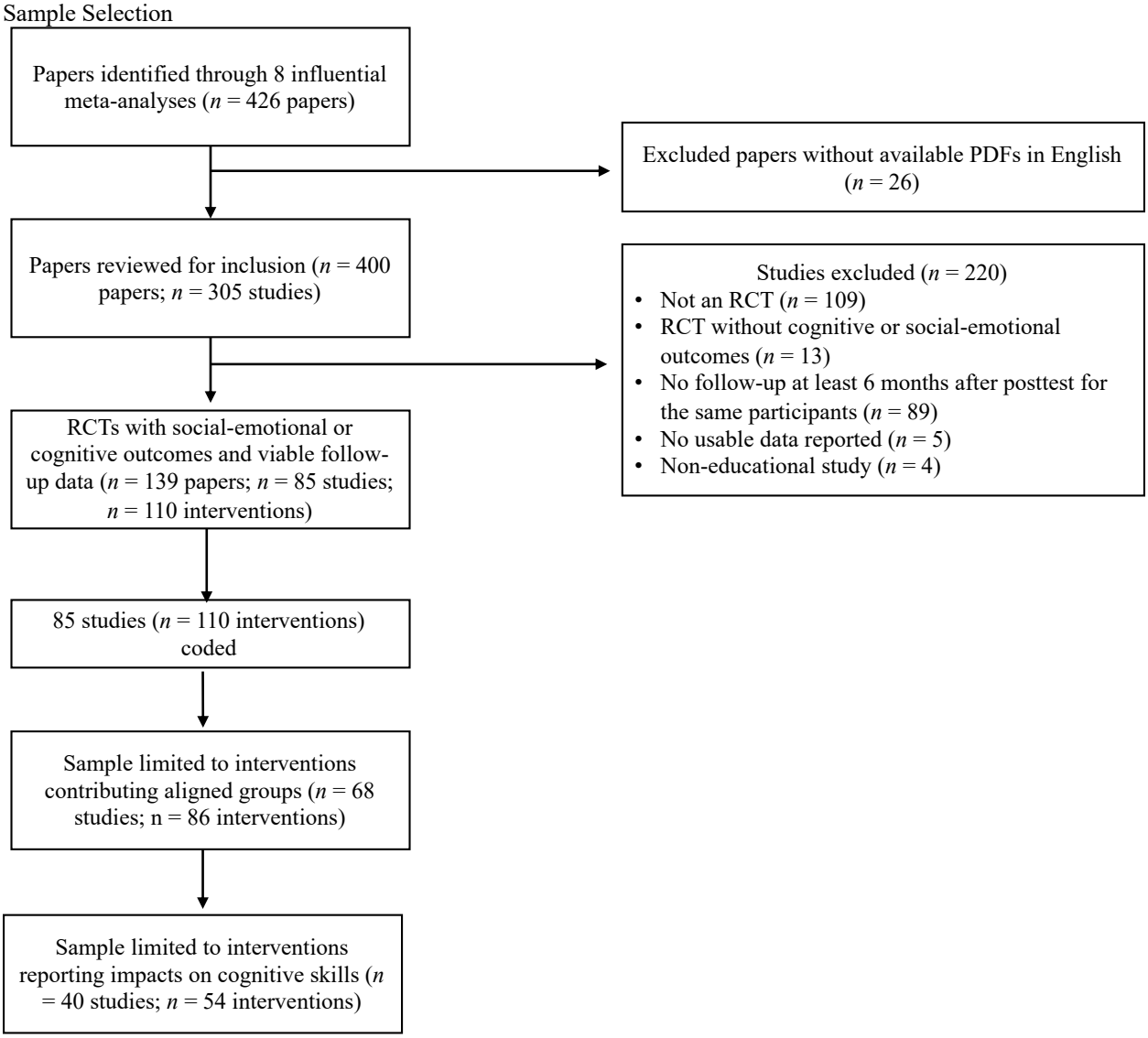
| | Confidence (1) | Intended Months of Treatment (2) | Age (years) (3) | Baseline Sample Size (4) |
|---|---|---|---|---|
| Intercept (null = 0.24 (0.04) ***) | 0.08 (0.04) | 0.07 (0.04) | 0.09 (0.04) * | 0.09 (0.04) * |
| Posttest | 0.49 (0.06) *** | 0.48 (0.04) *** | 0.45 (0.06) *** | 0.44 (0.04) ** |
| Unconstrained | -0.01 (0.02) | 0.02 (0.02) | -0.01 (0.02) | 0.00 (0.04) |
| Unconstrained * Posttest | -0.22 (0.09) * | -0.22 (0.10) | -0.08 (0.15) | -0.16 (0.06) |
| Confidence Level | 0.02 (0.01) | | | |
| Confidence Level * Posttest | 0.02 (0.02) | | | |
| Intervention Intended Months | | 0.02 (0.04) | | |
| Intervention Intended Months * Posttest | | -0.03 (0.03) | | |
| Baseline Age (years) | | | -0.01 (0.03) | |
| Baseline Age (years) * Posttest | | | 0.10 (0.10) | |
| Baseline Sample Size | | | | -0.02 (0.03) |
| Baseline Sample Size * Posttest | | | | -0.09 (0.03) |
| Random intercept (n = 0.16) | 0.11 | 0.11 | 0.10 | 0.10 |
| Random slope | 0.05 | 0.04 | 0.11 | 0.00 |
| $I^2$ (null = 52.98%) | 0.78% | 0.91% | -2.54% | -11.51% |
| *N* (Aligned groups/Treatments/Studies) | 153/43/31 | 175/43/31 | 175/43/31 | 175/43/31 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note. The above table shows models wherein covariates were interacted with posttest impacts. All models predict effect sizes at the 6–12-month follow-up. The unit of analysis is aligned groups of posttests and impacts at 6-to-12 months after the posttest wherein the construct, measure, subscale, and reporter were the same. Models were fit using the "metafor" package in R, with a study-level random effect, random slopes for posttest effect size, weighting, and robust standard errors. Standard errors are presented in parentheses, and coefficients can be interpreted in effect size units (i.e., the underlying data are in the effect size units reported by studies). The "null" model indicates a model estimating the average impact at the 6-12 month follow-up. Unconstrained refers to an indicator variable wherein "1" indicates an unconstrained aligned group and "0" indicates a constrained group. Heterogeneity statistics are shown for the null model and each of the subsequent models estimated above. Note that 9 unconstrained and 5 constrained aligned groups were missing values for treatment months. Missing values were replaced with the mean value of treatment months for each skill type (e.g., unconstrained aligned groups that were missing treatment months were assigned the

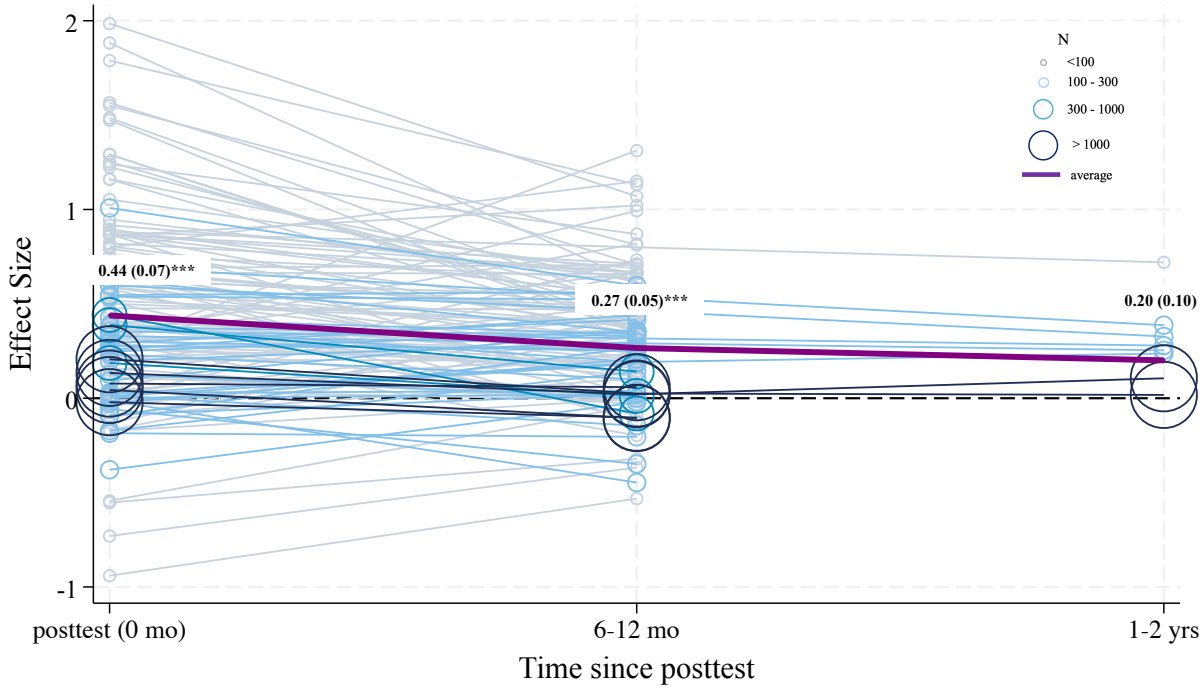
mean value of treatment months among other unconstrained aligned groups with data at the 6–12-month follow-up). None of the aligned groups were missing values for baseline age. Note that 24 aligned groups were coded in the initial coding effort as coders learned the coding scheme. These confidence ratings are therefore not included in the current analysis. Average confidence rating intended months of treatment, participant age, and baseline sample size were mean-centered and standardized. Note negative $I^2$ values are possible in meta-analyses with few studies, and might suggest a larger true value of heterogeneity (von Hippel, 2015)

Figure 1
Sample Selection



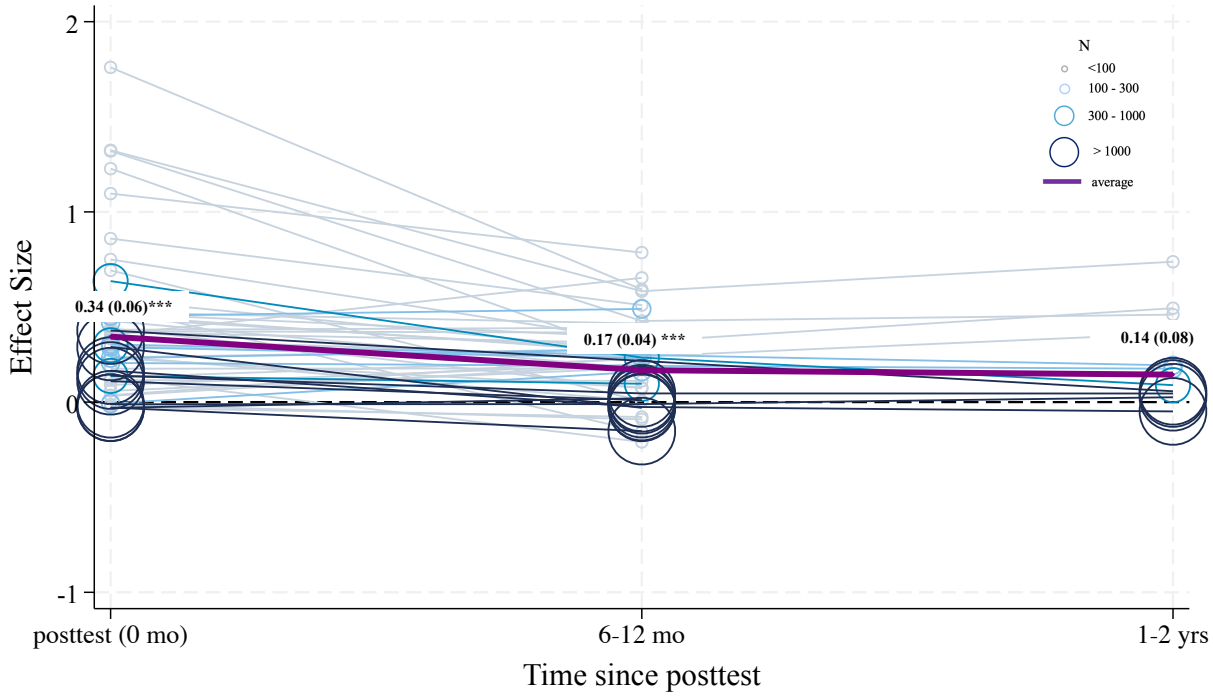

*Note*: Intervention refers to treatment-control group contrasts. Aligned groups refer to impacts measured using the same measure, subscale, construct, and reporter at posttest and at least one follow-up assessment.

Figure 2
Constrained Skills Fadeout Trajectories



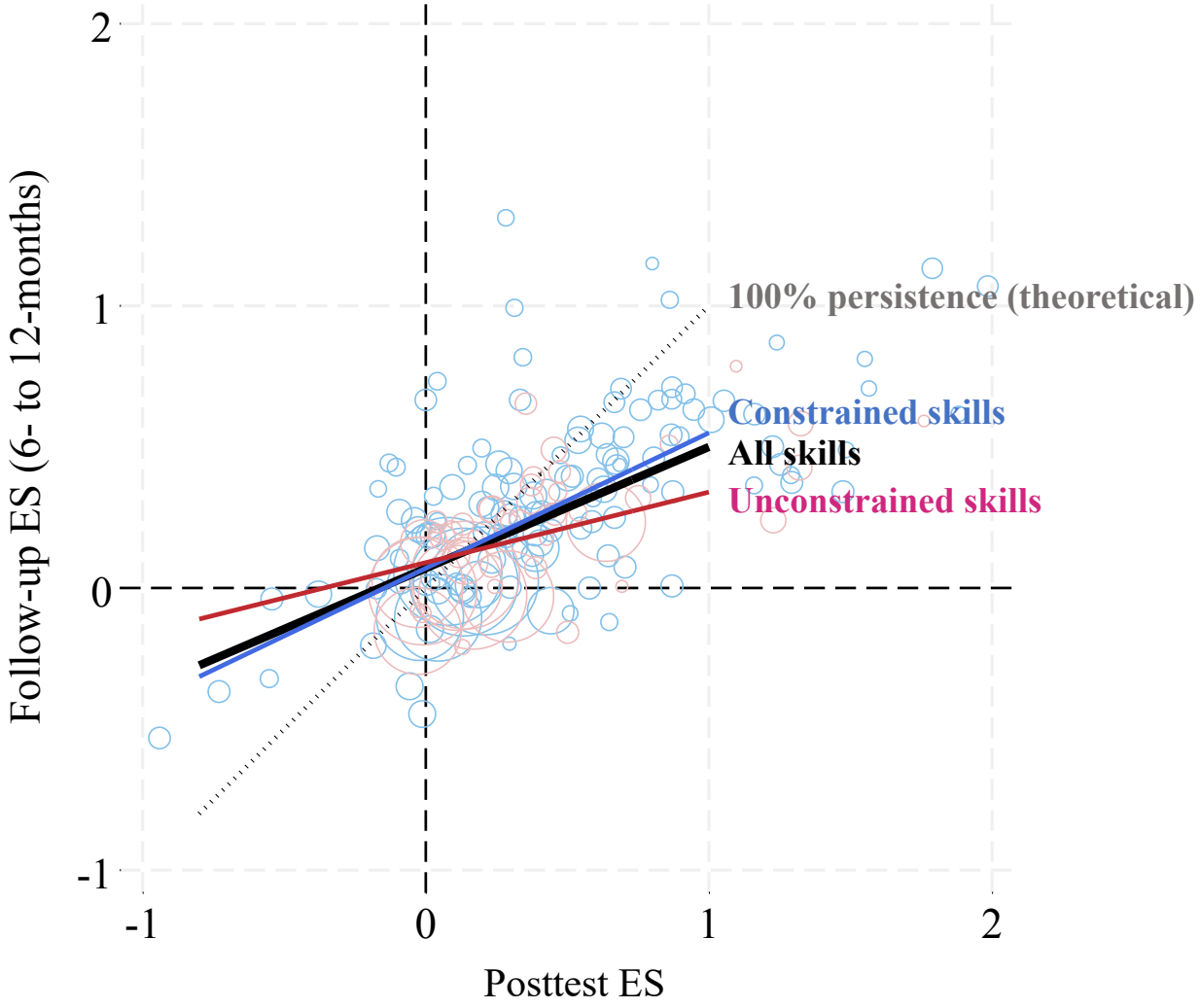

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
*Note*: Each line represents the treatment impact trajectory for a single constrained aligned group (i.e., construct was measured at posttest and at least one follow-up assessment using the same measure and subscale for the same intervention). The purple line displays the weighted meta-analytic average of effects at posttest, 6- to 12-months follow-up, and 1- to 2-years follow-up, calculated with a study-level random effect, random slopes for posttest effect size, weighting, and robust standard errors. Average effect sizes (in SDs) are shown for each coordinate with standard errors in parentheses. The dotted black line indicates an effect size of zero SD. Coordinates were weighted and color-coded according to sample size at posttest (larger circles and darker colors represent estimates from larger samples). Aligned groups with posttest effects below -1 standard deviations are not displayed for visual purposes ($n = 1$).

Figure 3
Unconstrained Skills Fadeout Trajectories




* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
*Note*: Each line represents the treatment impact trajectory for a single unconstrained aligned group (i.e., construct was measured at posttest and at least one follow-up assessment using the same measure and subscale for the same intervention). The purple line displays the weighted meta-analytic average of effects at posttest, 6- to 12-months follow-up, and 1- to 2-years follow-up, calculated with a study-level random effect, random slopes for posttest effect size, weighting, and robust standard errors. Average effect sizes (in SDs) are shown for each coordinate with standard errors in parentheses. The dotted black line indicates an effect size of zero SD. Coordinates were weighted and color-coded according to sample size at posttest (larger circles and darker colors represent estimates from larger samples).

Figure 4
Persistence Patterns at 6- to 12-Month Follow-up




*Note*: "ES" = effect size. The above figure displays persistence patterns based on estimates from meta-analytic regressions in which 6–12-month follow-up effect sizes were regressed on posttest effect sizes, an indicator for skill type, and the interaction between this indicator and posttest effect size (Table 4, Column 3). Models were fit using the "metafor" package in R, with a study-level random effect, random slopes for posttest effect size, weighting, and robust and clustered standard errors. The gray "100% persistence rate" line depicts what would be observed if posttest effect sizes perfectly predicted follow-up effect sizes (i.e., 100% conditional persistence), and no portion of the follow-up effect was unexplained by posttest effects, as indicated by the 0 intercept. The blue line shows the modeled persistence rate for constrained skills while the pink line shows modeled persistence rates for unconstrained skills. Only posttest and follow-up effects within the -1 to 2 SD range are presented for display purposes.

**Supplemental File for:**

# Using Meta-Analytic Data to Examine Fadeout and Persistence of Intervention Impacts on Constrained and Unconstrained Skills

Table S1

Codes Contributed by Sample Interventions

| Intervention | Constrained | Unconstrained | Exclude | Developmental Period | Intervention Type |
|---|---|---|---|---|---|
| Abecedarian~ All Pre-K (tx + tx; tx + cntrl) Group | | X | | Infancy/Toddlerhood | Early Childhood Education |
| Classroom and At-Home Preschool Interventions | | X | X | Early Childhood | Early Childhood Education |
| Classroom-Centered and School-Family Partnership Interventions~ Classroom-Centered Group | | | X | Middle childhood | Early Childhood Education |
| Classroom-Centered and School-Family Partnership Interventions~ Family-School Partnership Group | | | X | Middle childhood | Early Childhood Education |
| Code-Oriented Reading Instruction | X | X | | Early Childhood | Literacy Intervention |
| Cogmed Working Memory Training | | | X | Middle childhood | Executive Functioning Intervention |
| Computer-Assisted Blending Skill Training | X | | | Early Childhood | Literacy Intervention |
| Computer-Assisted Learning Program | X | | X | Middle childhood | Literacy Intervention |
| Computer-Assisted Reading Intervention for Children at Risk of Dyslexia~ Read, Write, and Type Group | X | X | | Middle childhood | Literacy Intervention |
| Computer-Assisted Reading Intervention for Children at Risk of Dyslexia~ The Lindamood Phoneme Sequencing Program for Reading, Spelling, and Speech Group | X | X | | Middle childhood | Literacy Intervention |
| Computer-Assisted Remedial Reading Intervention | X | | | Middle childhood | Literacy Intervention |
| Dialogic Reading #1~ School plus Home Reading Group | | X | | Early Childhood | Literacy Intervention |
| Dialogic Reading #1~ School Reading Group | | X | | Early Childhood | Literacy Intervention |
| Dialogic Reading #2 | | X | | Infancy/Toddlerhood | Literacy Intervention |

| Intervention | Constrained | Unconstrained | Exclude | Developmental Period | Intervention Type |
|---|---|---|---|---|---|
| English Reading Intervention for English Language Learners~ Study 1 & 4 | X | X | X | Middle childhood | Literacy Intervention |
| Explicit Phonological Awareness Instruction | X | | X | Middle childhood | Literacy Intervention |
| Head Start Classroom-based Approaches and Resources for Emotion and Social Skill Promotion~ Incredible Years Teacher Training Group | | X | X | Early Childhood | Socioemotional Learning |
| Head Start Classroom-based Approaches and Resources for Emotion and Social Skill Promotion~ Preschool Promoting Alternative Thinking Strategies Group | | X | X | Early Childhood | Socioemotional Learning |
| Head Start Classroom-based Approaches and Resources for Emotion and Social Skill Promotion~ Tools of the Mind Play Group | | X | X | Early Childhood | Socioemotional Learning |
| Head Start Impact Study | X | X | | Early Childhood | Early Childhood Education |
| Head Start Research-Based, Developmentally-Informed Program | X | X | | Early Childhood | Early Childhood Education |
| Home-Based Dyslexia Prevention | X | | | Early Childhood | Literacy Intervention |
| Infant Health and Development Program | | X | | Infancy/Toddlerhood | Early Childhood Education |
| Living Letters #2~ Combined Text Comprehension and Oral Language Group | X | X | | Middle childhood | Literacy Intervention |
| Living Letters #2~ Oral Language Group | X | X | | Middle childhood | Literacy Intervention |
| Living Letters #2~ Text Comprehension Group | X | X | | Middle childhood | Literacy Intervention |
| Multi-Component Reading Remediation~ Phonological Analysis and Blending/Direct | X | X | X | Middle childhood | Literacy Intervention |

| Intervention | Constrained | Unconstrained | Exclude | Developmental Period | Intervention Type |
|---|---|---|---|---|---|
| Instruction and Retrieval, Automaticity, Vocabulary, Engagement with Language, and Orthography Group | | | | | |
| Multi-Component Reading Remediation~ Phonological Analysis and Blending/Direct Instruction and Word Identification Strategy Training Group | X | X | X | Middle childhood | Literacy Intervention |
| Omega-Interactive Sentences, Computerized Phonological Training~ Combined Phonological and Comprehension Training Group | X | X | | Middle childhood | Literacy Intervention |
| Omega-Interactive Sentences, Computerized Phonological Training~ Comprehension Training Group | X | X | | Middle childhood | Literacy Intervention |
| Omega-Interactive Sentences, Computerized Phonological Training~ Phonological Training Group | X | X | | Middle childhood | Literacy Intervention |
| Parent Training for Teenage Moms | | | X | Infancy/Toddlerhood | Home Visiting |
| Perry Preschool | | X | X | Early Childhood | Early Childhood Education |
| Phonics-Based Instruction for First Graders | X | | X | Early Childhood | Literacy Intervention |
| Phonological/Early Reading Skills | X | | | Middle childhood | Literacy Intervention |
| Read Well Kindergarten | X | X | | Early Childhood | Literacy Intervention |
| Reading Recovery | X | | X | Early Childhood | Literacy Intervention |
| Reading Remediation | X | | X | Middle childhood | Literacy Intervention |
| Reading Remediation for Children with Reading Disorders | X | X | X | Middle childhood | Literacy Intervention |

| Intervention | Constrained | Unconstrained | Exclude | Developmental Period | Intervention Type |
|---|---|---|---|---|---|
| Reading with Rhyme, Reading with Phoneme~ Reading with Phoneme Group | X | | | Early Childhood | Literacy Intervention |
| Reading with Rhyme, Reading with Phoneme~ Reading with Rhyme and Phoneme Group | X | | | Early Childhood | Literacy Intervention |
| Reading with Rhyme, Reading with Phoneme~ Reading with Rhyme Group | X | | | Early Childhood | Literacy Intervention |
| SEARCH Screening Test and TEACH Tutoring~ Phonetic Group | X | | X | Early Childhood | Literacy Intervention |
| SEARCH Screening Test and TEACH Tutoring~ TEACH Tutoring Group | X | | X | Early Childhood | Literacy Intervention |
| Spanish Reading Intervention for English Language Learners~ Study 2 & 3 | X | X | | Middle childhood | Literacy Intervention |
| Supplemental Phonics-based Instruction | X | | | Early Childhood | Literacy Intervention |
| Supplemental Reading Instruction | X | X | | Middle childhood | Literacy Intervention |
| Swedish Phonics-based Intervention | X | | X | Middle childhood | Literacy Intervention |
| Teacher Responsivity Education | | X | | Early Childhood | Literacy intervention |
| Technology-Enhanced, Research-Based, Instruction, Assessment, and Professional Development Scale-up Model~ Building Blocks Group | | X | | Early Childhood | Math Intervention |
| Technology-Enhanced, Research-Based, Instruction, Assessment, and Professional Development Scale-up Model~ Building Blocks plus TRIAD Follow-Through Group | | X | | Early Childhood | Math Intervention |
| Tennessee Pre-K | | | X | Early Childhood | Early Childhood Education |

| Intervention | Constrained | Unconstrained | Exclude | Developmental Period | Intervention Type |
|---|---|---|---|---|---|
| The Early Training Project~ 3-year and 2-year Intervention Group | | X | X | Early Childhood | Early Childhood Education |
| Tools for Getting Along | | | X | Middle childhood | Socioemotional Learning |

*Note*: Interventions could contribute more than one code because interventions could contribute more than one aligned group measuring different skills. Interventions are listed by treatment-control group contrast. For instance, Living Letters #2 is shown as two separate interventions in the table above because it included two treatment groups. Developmental period refers to the developmental period at baseline. Interventions targeting students aged three and below were considered infancy/toddlerhood. Interventions targeting students between ages three and six were considered early childhood. Interventions targeting children between six and 11 were considered middle childhood. Intervention type refers to the broad focus of the intervention.

Table S2

Codes for Construct, Measure, and Subscale Combinations

| Construct | Measure | Subscale | Code |
|---|---|---|---|
| oral passage reading | analytic reading inventory | . | Constrained |
| letter word identification-spanish | batería r woodcock-muñoz dictado | identificación de letras y palabras | Constrained |
| context-free word recognition | burt word reading test | . | Constrained |
| word reading | burt word reading test | . | Constrained |
| number | cited measure (bas basic numbers skills test) | . | Constrained |
| rapid naming of letters | cited measure (denckla & rudel) | . | Constrained |
| spelling | cited measure (foorman) | . | Constrained |
| word reading | cited measure (foorman) | . | Constrained |
| oral reading fluency | cited measure (makar) | . | Constrained |
| spoonerism | cited measure (perin) | . | Constrained |
| phonological awareness | comprehensive test of phonological processing | blending phonemes-words, blending phonemes-non words, segmenting phonemes, and phoneme elison subtests | Constrained |
| phonological awareness-blending words | comprehensive test of phonological processing | blending words test | Constrained |
| elison | comprehensive test of phonological processing | elison subtest | Constrained |
| phonological awareness- elision | comprehensive test of phonological processing | elison test | Constrained |
| nonword repetition | comprehensive test of phonological processing | nonword repetition subtest | Constrained |
| rapid digit naming | comprehensive test of phonological processing | rapid digit naming test | Constrained |
| rapid letter naming | comprehensive test of phonological processing | rapid letter naming test | Constrained |
| phonological awareness-segmenting words | comprehensive test of phonological processing | segmenting words test | Constrained |

| Construct | Measure | Subscale | Code |
|---|---|---|---|
| common words recognition | daniels- diack oral reading tests | . | Constrained |
| consonant blends beginning recognition | daniels- diack oral reading tests | . | Constrained |
| consonant blends ending recognition | daniels- diack oral reading tests | . | Constrained |
| graded phonetically complex word recognition | daniels- diack oral reading tests | . | Constrained |
| nonsense syllables recognition | daniels- diack oral reading tests | . | Constrained |
| phonetically simple word recognition | daniels- diack oral reading tests | . | Constrained |
| polysyllabic- phonetically simple word recognition | daniels- diack oral reading tests | . | Constrained |
| reading quotient | daniels- diack oral reading tests | . | Constrained |
| reversible words recognition | daniels- diack oral reading tests | . | Constrained |
| spelling | dls | . | Constrained |
| reading fluency- english | dynamic indicators of basic early literacy skills | . | Constrained |
| nonword reading | graded nonword reading test | . | Constrained |
| reading fluency- spanish | indicadores dinámicos del exito en la lectura | . | Constrained |
| reading fluency | lukilasse graded fluency test | . | Constrained |
| word reading | multiple measures | . | Constrained |
| passage reading | passage reading test | . | Constrained |
| phonemic awareness | phonemic awareness test | . | Constrained |
| blend score | pollack tests | auditory tests | Constrained |
| phoneme discrimination | pollack tests | auditory tests | Constrained |
| basic reading | roswell-chall tests | auditory blending | Constrained |
| consonant combination | roswell-chall tests | diagnostic reading | Constrained |

| Construct | Measure | Subscale | Code |
|---|---|---|---|
| rule of silent e | roswell-chall tests | diagnostic reading | Constrained |
| short vowels | roswell-chall tests | diagnostic reading | Constrained |
| single consonant sounds | roswell-chall tests | diagnostic reading | Constrained |
| vowel combination | roswell-chall tests | diagnostic reading | Constrained |
| spelling | stanford achievement tests | . | Constrained |
| word study skills | stanford achievement tests | . | Constrained |
| blending | study-created measure | . | Constrained |
| letter identification | study-created measure | . | Constrained |
| non-word reading | study-created measure | . | Constrained |
| oral reading fluency | study-created measure | . | Constrained |
| phoneme deletion | study-created measure | . | Constrained |
| reading speed | study-created measure | . | Constrained |
| receptive letter knowledge | study-created measure | . | Constrained |
| reversed spoonerism | study-created measure | . | Constrained |
| segmentation | study-created measure | . | Constrained |
| spelling | study-created measure | . | Constrained |
| phoneme blending | test for phoneme blending | . | Constrained |
| phoneme segmentation | test for phoneme segmentation | . | Constrained |
| word reading efficiency | test of word reading efficiency | . | Constrained |
| word reading efficiency nonwords | test of word reading efficiency | . | Constrained |
| word reading efficiency real words | test of word reading efficiency | . | Constrained |
| reading efficacy | test of word reading efficiency | phonemic decoding and sight word efficiency subtests | Constrained |
| phonemic decoding efficiency | test of word reading efficiency | phonemic decoding efficiency subtest | Constrained |

| Construct | Measure | Subscale | Code |
|---|---|---|---|
| blending/phoneme decoding | test of word reading efficiency | phonemic decoding subscale | Constrained |
| elison/sight word reading | test of word reading efficiency | sight word efficiency subscale | Constrained |
| word efficiency | test of word reading efficiency | word efficiency subtest | Constrained |
| sight word reading | test of word reading efficiency | word subtest | Constrained |
| word recognition | time2 | . | Constrained |
| segment subtraction | umesol | segment subtraction subtest | Constrained |
| spelling | waddington diagnostic spelling test | . | Constrained |
| arithmetic | wechsler | numerical operations | Constrained |
| spelling | wide-range achievement test | spelling | Constrained |
| developmental spelling | wide-range achievement test | spelling subtest | Constrained |
| spelling | wide-range achievement test | spelling subtest | Constrained |
| math applied problems | woodcock johnson | applied problems subtest | Constrained |
| letter name identification | woodcock johnson | letter-word id | Constrained |
| pre-academic skills | woodcock johnson | pre-academic skills composite (letter-word identification, spelling, applied problems) | Constrained |
| word attack | woodcock johnson | word attack | Constrained |
| word attack | woodcock johnson | word attack subtest | Constrained |
| print awareness/letter word | woodcock johnson | word identification subscale | Constrained |
| word attack- english | woodcock language proficiency battery | word attack subtest | Constrained |
| word attack- spanish | woodcock language proficiency battery | word attack subtest | Constrained |
| basic skills cluster | woodcock reading mastery test | basic skills cluster | Constrained |
| word analysis | woodcock reading mastery test | word attack subtest | Constrained |
| word attack | woodcock reading mastery test | word attack subtest | Constrained |
| word id | woodcock reading mastery test | word id subtest | Constrained |
| reading accuracy | woodcock reading mastery test | word identification and word attack subtests | Constrained |
| word reading | woodcock reading mastery test | word identification and word attack subtests | Constrained |

| Construct | Measure | Subscale | Code |
|---|---|---|---|
| word id | woodcock reading mastery test | word identification subtest | Constrained |
| word identification | woodcock reading mastery test | word identification subtest | Constrained |
| phonological recoding | word attack skills test | . | Constrained |
| word recognition | word-chains test | . | Constrained |
| general knowledge | academic rating scale | . | Unconstrained |
| expressive vocabulary | clinical evaluation of language fundamentals | expressive vocabulary subtest | Unconstrained |
| expressive vocabulary | expressive one-word picture vocabulary test | . | Unconstrained |
| vocabulary | expressive one-word picture vocabulary test | . | Unconstrained |
| auditory-vocal association | illinois test of psycholinguistic abilities | auditory-vocal association | Unconstrained |
| verbal expressiveness | illinois test of psycholinguistic abilities | expressive subscale | Unconstrained |
| verbal fluency | illinois test of psycholinguistic abilities | expressive subscale | Unconstrained |
| language deficiencies | illinois test of psycholinguistic abilities | total language age | Unconstrained |
| reading comprehension | neale analysis of reading ability | reading comprehension | Unconstrained |
| receptive language | peabody picture vocabulary test | . | Unconstrained |
| receptive vocabulary | peabody picture vocabulary test | . | Unconstrained |
| vocabulary | peabody picture vocabulary test | . | Unconstrained |
| math achievement | rema | . | Unconstrained |
| reading comprehension | stanford achievement tests | . | Unconstrained |
| vocabulary | stanford achievement tests | . | Unconstrained |
| iq | stanford-binet intelligence scale | . | Unconstrained |
| receptive vocabulary- spanish | test de vocabulario en imagenes peabody | . | Unconstrained |
| iq | wechsler | . | Unconstrained |
| reading comprehension | wechsler | reading comprehension | Unconstrained |
| vocabulary | wechsler | vocabulary | Unconstrained |
| comprehension | woodcock johnson | comprehension | Unconstrained |
| oral comprehension | woodcock johnson | oral comprehension | Unconstrained |

| Construct | Measure | Subscale | Code |
|---|---|---|---|
| vocabulary | woodcock johnson | vocabulary | Unconstrained |
| passage comprehension-english | woodcock language proficiency battery | passage comprehension subtest | Unconstrained |
| passage comprehension-spanish | woodcock language proficiency battery | passage comprehension subtest | Unconstrained |
| reading comprehension | woodcock reading mastery test | . | Unconstrained |
| comprehension | woodcock reading mastery test | passage comprehension | Unconstrained |
| passage comprehension | woodcock reading mastery test | passage comprehension subtest | Unconstrained |
| passage comprehension | woodcock reading mastery test | passive comprehension subtest | Unconstrained |
| language and literacy | academic rating scale | . | Exclude |
| mathematical thinking | academic rating scale | . | Exclude |
| verbal working memory | automated working memory assessment | backward digit recall subtest | Exclude |
| verbal short term memory | automated working memory assessment | digit recall subtest | Exclude |
| visuospatial short term memory | automated working memory assessment | dot matrix subtest | Exclude |
| visuospatial working memory | automated working memory assessment | mister x score subtest | Exclude |
| word reading- irregular words | batterie d'evaluation du langage ecrit | irregular words | Exclude |
| mental score | bayley mental development index | . | Exclude |
| metacognition | behavior rating inventory of executive function | metacognition index | Exclude |
| book level | clay's book level test | . | Exclude |
| math achievement | comprehensive test of basic skills | . | Exclude |
| reading achievement | comprehensive test of basic skills | . | Exclude |
| reading age | daniels- diack oral reading tests | . | Exclude |
| visual perception | frostig mean pq | . | Exclude |
| oral reading quotient | gray oral reading test | . | Exclude |
| reading ability | gray oral reading test | . | Exclude |
| oral reading quotient | gray oral reading test | overall quotient | Exclude |

| Construct | Measure | Subscale | Code |
|---|---|---|---|
| non-verbal ability | leiter international performance scale | . | Exclude |
| reading achievement | multiple measures | . | Exclude |
| iq | na | . | Exclude |
| word decoding | na | . | Exclude |
| reading ability | neale analysis of reading ability | . | Exclude |
| word recognition accuracy | neale analysis of reading ability | accuracy subtest | Exclude |
| syllabification | roswell-chall tests | diagnostic reading | Exclude |
| language | stanford achievement tests | . | Exclude |
| math application | stanford achievement tests | . | Exclude |
| math computation | stanford achievement tests | . | Exclude |
| math concepts | stanford achievement tests | . | Exclude |
| science | stanford achievement tests | . | Exclude |
| social studies | stanford achievement tests | . | Exclude |
| arithmetic | wide-range achievement test | . | Exclude |
| reading | wide-range achievement test | . | Exclude |
| reading ability | wide-range achievement test | . | Exclude |
| spelling | wide-range achievement test | . | Exclude |
| math calculations | woodcock johnson | calculations subtest | Exclude |
| cognitive achievement | woodcock johnson | letter-word identification, spelling, oral comprehension, picture vocabulary, applied problems, quantitative concepts | Exclude |
| oral language composite-english | woodcock language proficiency battery | . | Exclude |
| oral language composite-spanish | woodcock language proficiency battery | . | Exclude |

| Construct | Measure | Subscale | Code |
|---|---|---|---|
| word reading | woodcock reading mastery test | average of word attack and word identification subtests | Exclude |

Table S3

Measures and Aligned Groups Excluded from the Primary Analyses

| Exclusion reason | Number of combinations of Construct, Measure, and Subscale | Number of aligned groups |
|---|---|---|
| Binary scale | 2 | 6 |
| Insufficient measure information | 6 | 6 |
| Multiple subscales | 25 | 34 |
| Unrelated to theory | 6 | 6 |

*Note*: The above table indicates the number of combinations of construct, measures, and subscales excluded from the primary analyses. “Aligned groups” refers to the groupings that included a posttest and at least one follow-up assessment of the same construct measured using the same measure, subscales, and reporter, within a treatment-control contrast. In total, 39 combinations of construct, measure, and subscale mapping onto 52 aligned groups were excluded from the main analyses.

Table S4

Combinations of Construct, Measure, and Subscale for which Two of Three Coders Agreed on a Code

| Construct | Measure | Subscale | Code for which Two Coders Agreed |
|---|---|---|---|
| mathematical thinking | Academic Rating Scale | . | Constrained |
| verbal working memory | Automated Working Memory Assessment | backward digit recall subtest | Unconstrained |
| visuospatial working memory | Automated Working Memory Assessment | mister x score subtest | Unconstrained |
| verbal short term memory | Automated Working Memory Assessment | digit recall subtest | Unconstrained |
| visuospatial short term memory | Automated Working Memory Assessment | dot matrix subtest | Unconstrained |
| word reading- irregular words | Batterie d'Evaluation du Langage Écrit | irregular words | Constrained |
| metacognition | Behavior Rating Inventory of Executive Function | metacognition index | Unconstrained |
| non-verbal ability | Leiter International Performance Scale | . | Unconstrained |
| word recognition accuracy | Neale Analysis of Word Reading Ability | accuracy subtest | Constrained |
| syllabification | Roswell-Chall Tests | diagnostic reading | Constrained |
| math computation | Stanford Achievement Tests | . | Constrained |
| oral language composite-spanish | Woodcock Language Proficiency Battery | . | Unconstrained |

*Note*: The current table shows the 12 unique combinations of construct, measure, and subscale for which two of the three coders agreed on a code. The 15 viable groups associated with these combinations were not included in the primary analysis but were included in a sensitivity analysis.

Table S5

Intervention and Participant Characteristics (mean [minimum- maximum])

| | **All outcomes (1)** | | **Excluded outcomes (2)** | |
|---|---|---|---|---|
| | **Mean** | ***N*** | **Mean** | ***N*** |
| Base sample size | 375.41 [24 - 3933] | 54 | 400.50 [24 -1323] | 24 |
| Publication year | 2005 [1969 - 2022] | 54 | 2002.30 [1969 - 2017] | 24 |
| Baseline age (months) | 67.80 [0.00 - 122.00] | 54 | 72.42 [0.00 - 122] | 24 |
| Intended months of treatment | 8.12 [0.92 - 36.00] | 46 | 7.86 [1.15 - 18.00] | 21 |
| Intended hours of treatment | 115.22 [4.00 - 1075.39] | 37 | 212.59 | 16 |
| Included extra time in school (%) | 23.40 | 47 | 19.05 | 21 |
| **Treatment Focus** | | | | |
| Math | 11.11 | 54 | 12.50 | 24 |
| Reading/Language | 83.33 | 54 | 70.83 | 24 |
| Science | 1.85 | 54 | 4.17 | 24 |
| General cognitive | 9.26 | 54 | 8.33 | 24 |
| Executive functioning | 1.85 | 54 | 4.17 | 24 |
| Learning skills | 1.85 | 54 | 4.17 | 24 |
| Social-emotional skills | 24.07 | 54 | 33.33 | 24 |
| Substance use | 0.00 | 54 | 0.00 | 24 |
| **Adult involvement (%)** | | | | |
| Teacher | 55.56 | 54 | 50.00 | 24 |
| Parent | 22.22 | 54 | 16.67 | 24 |
| **Race/Ethnicity (%)** | | | | |
| Asian | 19.08 | 10 | 17.60 | 4 |
| Black | 46.99 | 30 | 48.14 | 18 |
| White | 42.85 | 22 | 53.33 | 13 |
| Hispanic | 33.08 | 20 | 33.97 | 9 |
| Female participants (%) | 44.85 | 44 | 45.24 | 21 |

*Note*: This table presents the intervention characteristics for unique treatment-control group contrasts that contributed to any of the coding categories. Column 1 shows characteristics for all outcomes contributing aligned groups to the analytic sample (constrained, unconstrained, and excluded). Column 2 shows characteristics of interventions contributing at least one treatment-control group contrast to an "excluded" aligned group. “N” indicates the number of treatment-control group contrasts that reported information and the percentage of treatment-control group contrasts that contributed to the averages. Note that 18 combinations of construct, measure, and subscale were coded as part of an initial coding round as coders learned the coding scheme. These 18 combinations

came from seven interventions (one of which contributed a constrained aligned groups and all of which contributed excluded aligned groups).

Table S6

Analytic Sample Characteristics for Excluded Aligned Groups

| | **Treatment groups (#) (1)** | **Aligned groupings (#) (2)** | **Avg # follow-ups (3)** | **Avg Posttest ES, weighted (SE) (4)** | **Avg Posttest ES, unweighted (SE) (5)** |
|---|---|---|---|---|---|
| **Excluded** | **24** | **52** | **1.23** | **0.15 (0.06) *** | **0.14 (0.12)** |
| Language and Literacy | 17 | 28 | 1.04 | 0.23 (0.09) * | 0.26 (0.13) |
| Cognitive | 6 | 9 | 1.89 | 0.14 (0.02) | 0.17 (0.08) |
| Math | 9 | 12 | 1.00 | -0.02 (0.11) | -0.11 (0.18) |
| Achievement Composite | 1 | 1 | 4.00 | | 0.32 (0.07) ❑ |
| Other Academic Ability | 1 | 2 | 1.00 | -0.17 (0.17) ° | -0.17 (0.17) ° |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

*Note*: "ES" = Effect size. The above table shows sample characteristics for aligned groups excluded from the main analyses for the reasons outlined in Table S3 (these aligned groups were not coded as either constrained or unconstrained). The number of aligned groups refers to the number of groupings that included a posttest and at least one follow-up assessment of the same construct measured using the same measure, subscales, and reporter within a treatment-control contrast. Average number of follow-ups reflects the average number of follow-up assessments that were collected (at least 6 months after the posttest) for each aligned group. Models were estimated using the "metafor" package in R. Weighted models included a study-level random effect, random slopes for posttest effect size, weighting, and robust standard errors. Unweighted effects were estimated using a fixed-effects meta-analytic model with no random or econometric fixed effect for study ID. Standard errors are presented in parentheses, and coefficients can be interpreted in effect size units (i.e., the underlying data are in the effect size units reported by studies). Effect sizes listed as "N/A" indicate there were insufficient aligned groups to generate an estimate.

° indicates cluster-robust standard errors could not be produced because all outcomes came from one study.

❑ indicates that only one aligned group contributed an outcome and we could not determine a weighted or unweighted average. The estimate shown in column 5 is the effect size and standard error for the one aligned group.

Table S7

Average Unweighted Meta-Analytic Effect Sizes at Each Binned Study Wave

| | **Constrained or Unconstrained (1)** | | **Constrained (2)** | | **Unconstrained (3)** | |
|---|---|---|---|---|---|---|
| | **Avg ES (SE)** | ***n*** | **Avg ES (SE)** | ***n*** | **Avg ES (SE)** | ***n*** |
| > 1 year, up to 2 years | 0.26 (0.10) | 19 | 0.29 (0.08) * | 8 | 0.24 (0.16) | 11 |
| >2 years, up to 3 years | 0.22 (0.07) | 6 | N/A | 0 | 0.22 (0.07) | 6 |
| > 3 years, up to 4 years | 0.21 (0.03) * | 18 | 0.23 (0.02) * | 8 | 0.19 (0.04) | 10 |
| > 4 years | 0.22 (0.08) | 11 | 0.26 (0.1) | 7 | 0.15 (0.09) | 4 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

*Note*: "ES" = effect size. Effect sizes are in standard deviation units. The unit of analysis is aligned groups of posttest and impacts at 6-to-12 months after the posttest wherein the construct, measure, subscale and reporter were the same. Column 1 shows the meta-analytic average effect size for all aligned groups coded as either constrained or unconstrained. Columns 2 and 3 show the meta-analytic average effect size at each assessment period for aligned groups coded as constrained and unconstrained, respectively. Each column shows the number of aligned groups at each wave ("*n*"). "N/A" indicates there were insufficient aligned groups to determine a weighted estimate. Models were fit using the "metafor" package in R and included a fixed-effects meta-analytic model with no random or econometric fixed effect for study ID. Standard errors are presented in parentheses.

Table S8

Ratio of posttest to follow-up standard deviations

| | Constrained or Unconstrained | | | | | Constrained | | | | | Unconstrained | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | *M* | *SD* | Min | Max | *n* | *M* | *SD* | Min | Max | *n* | *M* | *SD* | Min | Max |
| **Control group** | | | | | | | | | | | | | | | |
| Ratio of posttest SD to 6–12-month FU SD | 131 | 1.23 | 0.61 | 0.38 | 6.45 | 97 | 1.24 | 0.70 | 0.38 | 6.45 | 34 | 1.21 | 0.24 | 0.87 | 2.06 |
| **Treatment** | | | | | | | | | | | | | | | |
| Ratio of posttest SD to 6–12-month FU SD | 131 | 1.14 | 0.43 | 0.25 | 4.33 | 97 | 1.17 | 0.49 | 0.25 | 4.33 | 34 | 1.08 | 0.20 | 0.70 | 1.56 |

Note: “SD” is standard deviation. “FU” is follow-up. The above table shows the ratio of posttest to 6- 12-month follow-up standard deviations for control and treatment groups. The current sample includes aligned groups for which papers (1) reported raw mean and standard deviation data, (2) score type did not change between papers or timepoints (from standardized to raw and vice versa), and (3) raw means were not determined by the coding team by averaging across papers or treatment groups. In total, 137 aligned groups met these criteria.

Table S9

Persistence Rates for Unconstrained & Constrained Outcomes at 1 to 2 Years Follow-Up

| | **Baseline** | **Skill Type** | **Interaction** |
|---|---|---|---|
| | **(1)** | **(2)** | **(3)** |
| Intercept (null = 0.19 (0.07) *) | 0.01 (0.01) | 0.05 (0.04) | -0.02 (0.02) |
| Posttest | 0.29 (0.1) | 0.28 (0.1) | 0.55 (0.15) * |
| Unconstrained | | -0.06 (0.05) | 0.00 (0.02) |
| Unconstrained * Posttest | | | -0.30 (0.15) |
| Random intercept (null = 0.16) | 0.11 | 0.11 | 0.11 |
| Random slope | 0.14 | 0.17 | 0.18 |
| $I^2$ (null = 43.48%) | 6.08% | 0.21% | -0.43 |
| *N* (Aligned groups/Treatments/Studies) | 19/8/8 | 19/8/8 | 19/8/8 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

*Note*: The unit of analysis is aligned groups of posttests and impacts at 1-2 years after the posttest wherein the construct, measure, subscale, and reporter were the same. The "null" model indicates a model estimating the average impact at the 1-to-2-year follow-up. Column 1 shows the model with the posttest effect added. Column 2 shows the model with the skill type dummy. Column 3 shows the primary model, which includes the interaction between the posttest and the unconstrained dummy. Unconstrained refers to a dummy wherein 1 indicates an unconstrained aligned group and 0 indicates a constrained group. Coefficients can be interpreted in effect size units (i.e., the underlying data are in the effect size units reported by studies). Models were executed using the "metafor" package in R and included a study-level random effect, random slopes for posttest effect size, weighting, and robust standard errors (shown in parentheses). Note negative $I^2$ values are possible in meta-analyses with few studies, and might suggest a larger true value of heterogeneity (von Hippel, 2015).

Table S10

Pairwise Correlations Between Effect Sizes, Growth, and Covariates

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| (1) Posttest ES | - | | | | | | |
| (2) 6–12-month follow-up ES | 0.66*** | - | | | | | |
| (3) Treatment group growth | 0.22** | 0.08 | - | | | | |
| (4) Control group growth | 0.32*** | -0.00 | 0.97*** | - | | | |
| (5) Baseline age (years) | 0.03 | 0.06 | -0.13 | -0.16 | - | | |
| (6) Treatment length (months) | -0.06 | -0.07 | -0.20* | -0.24** | -0.40*** | - | |
| (7) Mean confidence rating | -0.05 | 0.08 | -0.10 | -0.10 | -0.02 | -0.15 | - |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

*Note:* "ES" indicates effect size. The above tables show pairwise correlations between effect sizes at posttest and the 6-to-12-month follow-up, control and treatment group growth, study characteristics, and average raters' average confidence in their code. Correlations are shown for aligned groups coded as either constrained or unconstrained (not excluded). Of the 184 aligned groups coded as constrained or unconstrained, 175 aligned groups had data at the posttest and at the 6 to 12-month follow-up. Growth refers to change in effect size from posttest to the 6 – 12-month follow-up and was generated as the follow-up minus the posttest score divided by the control group's posttest standard deviation for both control and treatment groups. Control and treatment group growth was determined only if (1) aligned groups reported raw data on means and standard deviations, (2) aligned groups were determined not to be reported in standardized units, and (3) the coding team did not determine raw means and standard deviations by averaging across papers reporting different means and standard deviations at the same time points. Thus, treatment and control group data are shown for 138 aligned group. One-hundred eighty-four aligned groups had data on baseline age. Six interventions did not indicate treatment length; thus 169 aligned groups include data on this variable. Finally, 24 aligned groups were coded in the initial coding effort as coders learned the coding scheme. The above table therefore reflects the mean confidence rating from 160 aligned groups.

Table S11

Sensitivity Test: Econometric Fixed Effects

| | Baseline | Skill Type | Interaction |
|---|---|---|---|
| | (1) | (2) | (3) |
| Posttest | 0.45 (0.05) * | 0.45 (0.05) * | 0.47 (0.04) ** |
| Unconstrained | | -0.01 (0.02) | 0.04 (0.04) |
| Unconstrained * Posttest | | | -0.29 (0.14) |
| *N* (Aligned groups/Treatments/Studies) | 175/43/31 | 175/43/31 | 175/43/31 |

* p < 0.05, ** p < 0.01, *** p < 0.001

*Note*: The unit of analysis is aligned groups of posttests and impacts at 6-to-12 months after the posttest wherein the construct, measure, subscale, and reporter were the same. Column 1 presents the results of an econometric fixed effects model with a fixed effect for study and weighting. Column 2 includes a covariate for skill type (either constrained or unconstrained). Finally, column 3 includes the interaction between skill type and posttest effect size.

Table S12

Sensitivity Test: Correlations and Hierarchical Effects Estimate

| | Interaction |
|---|---|
| Intercept | 0.02 (0.03) |
| Posttest | 0.48 (0.03) *** |
| Unconstrained | 0.01 (0.04) |
| Unconstrained * Posttest | -0.17 (0.10) |
| *N* (Aligned groups/Treatments/Studies) | 175/43/31 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

*Note*: The current model used the Correlation and Hierarchical Effects (CHE; Pustejovsky & Tipton, 2022) and assumed a constant correlation coefficient of 0.60. The CHE Model allows for within and between study heterogeneity in effect sizes. We assumed a nested random effects model wherein aligned groups are nested within treatments which are nested within studies.

Table S13

Sensitivity Test: Only Including Measures with a Confidence Rating of 3

| | Baseline | Skill Type | Interaction |
|---|---|---|---|
| | (1) | (2) | (3) |
| Intercept (null = 0.26 (0.05) ***) | 0.07 (0.04) | 0.09 (0.04) | 0.08 (0.04) |
| Posttest | 0.48 (0.04) *** | 0.48 (0.04) *** | 0.52 (0.05) *** |
| Unconstrained | | -0.06 (0.03) | -0.03 (0.04) |
| Unconstrained * Posttest | | | -0.17 (0.07) |
| Random intercept (null = 0.00) | 0.09 | 0.09 | 0.09 |
| Random slope | 0.01 | 0.01 | 0.01 |
| $I^2$ (null = 49.25%) | -27.68% | -39.96% | 40.53 |
| *N* (Aligned groups/Treatments/Studies) | 99/32/24 | 99/32/24 | 99/32/24 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

*Note*. The unit of analysis is aligned groups of posttests and impacts at 6-to-12 months after the posttest wherein the construct, measure, subscale, and reporter were the same. The current analysis is limited to aligned groups for which the confidence rating for the constrained/unconstrained ratings was three on average between the three coders, indicating all coders were highly confident in their code. Seventy-four constrained and 25 unconstrained aligned groups had a mean confidence rating of three. Models were executed using the "metafor" package in R and included a study-level random effect, random slopes for posttest effect size, weighting, and robust standard errors. Standard errors are presented in parentheses, and coefficients can be interpreted in effect size units (i.e., the underlying data are in the effect size units reported by studies). Unconstrained refers to an indicator variable wherein "1" indicates an unconstrained aligned group and "0" indicates a constrained group. The null model indicates a model estimating the average impact at the 6-12 month follow-up. Columns 1 through 3 shows the model with just posttest, the dummy variable for skill type, and the interaction, respectively. Heterogeneity statistics are shown for the null model and each of the subsequent models estimated above. Note negative $I^2$ values are possible in meta-analyses with few studies, and might suggest a larger true value of heterogeneity (von Hippel, 2015).

Table S14

Sensitivity Test: Including Previously Excluded Codes

| | Baseline | Skill Type | Interaction |
|---|---|---|---|
| | (1) | (2) | (3) |
| Intercept (null = 0.23 (0.04) ***) | 0.06 (0.03) | 0.07 (0.04) | 0.06 (0.04) |
| Posttest | 0.47 (0.07) *** | 0.47 (0.07) *** | 0.5 (0.08) *** |
| Unconstrained | | -0.02 (0.02) | -0.01 (0.02) |
| Unconstrained * Posttest | | | -0.09 (0.11) |
| Random intercept (null = 0.00) | 0.10 | 0.10 | 0.10 |
| Random slope | 0.21 | 0.22 | 0.22 |
| $I^2$ (null = 55.22%) | 11.11% | 10.49% | 10.23% |
| *N* (Aligned groups/Treatments/Studies) | 188/50/37 | 188/50/37 | 188/50/37 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

*Note*: The unit of analysis is aligned groups of posttest and the 6-12 months follow-up impact collected for the same construct using the same measure, subscale, and reported at posttest and at follow-up. In the above analysis we included the codes for which only two of the three coders agree on a code. These codes were excluded from the main analyses. The inclusion of these previously excluded codes resulted in 133 aligned groups coded as constrained and 55 aligned groups coded as unconstrained. The "null" model indicates a model estimating the average impact at the 6 – 12 month follow-up. Column 1 shows the results when the 6-12 months follow-up effect size is regressed on the posttest effect size. Column 2 is the same as the column 1 model with a covariate for skill type (constrained or unconstrained). Column 3 adds an interaction between skill type and posttest effect size. Unconstrained refers to an indicator variable wherein 1 indicates an unconstrained aligned group and 0 indicates a constrained group. Models were fit using the "metafor" package in R, with a study-level random effect, random slopes for posttest effect size, weighting, and robust standard errors. Standard errors are presented in parentheses, and coefficients can be interpreted in effect size units (i.e., the underlying data are in the effect size units reported by studies). Heterogeneity statistics are presented for the null and subsequent models.

Table S15

Sensitivity Test: Language and Literacy Outcomes Only

| | Baseline | Skill Type | Interaction |
|---|---|---|---|
| | (1) | (2) | (3) |
| Intercept (null = 0.27 (0.04) ***) | 0.07 (0.04) | 0.09 (0.04) * | 0.09 (0.04) * |
| Posttest | 0.45 (0.04) ** | 0.44 (0.04) ** | 0.47 (0.03) ** |
| Unconstrained | | -0.04 (0.02) | -0.01 (0.02) |
| Unconstrained * Posttest | | | -0.16 (0.06) * |
| Random intercept (null = 0.00) | 0.12 | 0.12 | 0.12 |
| Random slope | 0.00 | 0.01 | 0.01 |
| $I^2$ (null = 52.92%) | 1.64% | -0.52% | -0.48% |
| *N* (Aligned groups/Treatments/Studies) | 161/29/22 | 161/29/22 | 161/29/22 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: The unit of analysis is aligned groups of posttest and the 6-12 months follow-up impact collected for the same construct using the same measure, subscale, and reported at posttest and at follow-up. For the sake of the current analysis, we limited the sample to aligned groups with measures of language or literacy. Limiting the sample to language/literacy measures yielded a sample of 119 constrained aligned groups and 42 unconstrained aligned groups. The "null" model indicates a model estimating the average impact at the 6-12 month follow-up. Column 1 shows the results when the 6-12 months follow-up effect size is regressed on the posttest effect size. Column 2 is the same as the column 1 model with a covariate for skill type (constrained or unconstrained). Column 3 adds an interaction between skill type and posttest effect size. Models were fit using the "metafor' package in R, with a study-level random effect, random slopes for posttest effect size, weighting, and robust standard errors (shown in parentheses). Heterogeneity statistics are presented for the null model and the subsequent three models. Note negative $I^2$ values are possible in meta-analyses with few studies, and might suggest a larger true value of heterogeneity (von Hippel, 2015).

Table S16

Sensitivity Test: Limiting to Early Childhood Interventions Only

| | Baseline | Skill Type | Interaction |
|---|---|---|---|
| | (1) | (2) | (3) |
| Intercept (null = 0.23 (0.04) ***) | 0.10 (0.03) ** | 0.12 (0.04) * | 0.11 (0.04) * |
| Posttest | 0.31 (0.05) *** | 0.31 (0.05) *** | 0.36 (0.07) ** |
| Unconstrained | | -0.04 (0.03) | -0.02 (0.02) |
| Unconstrained * Posttest | | | -0.12 (0.09) |
| Random intercept (null = 0.00) | 0.09 | 0.09 | 0.09 |
| Random slope | 0.07 | 0.08 | 0.08 |
| $I^2$ (null = 42.78%) | 9.85% | 4.38% | 4.92 |
| *N* (Aligned groups/Treatments/Studies) | 114/34/25 | 114/34/25 | 114/34/25 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

*Note*: The unit of analysis is aligned groups of posttest 6-12 months follow-up impact collected for the same construct using the same measure, subscale, and reported at posttest and at follow-up. For the sake of the current analysis, we limited the sample to aligned groups from studies in which the sample was aged seven years or below at baseline ($n$ = 114 aligned groups). Seventy-nine aligned groups contributed constrained outcomes and 35 aligned groups contributed unconstrained outcomes to the current analysis. The "null" model indicates a model estimating the average impact at the 6 – 12 month follow-up. Column 1 shows the results when the 6-12 months follow-up effect size is regressed on the posttest effect size. Column 2 is the same as the column 1 model with a covariate for skill type (constrained or unconstrained). Column 3 adds an interaction between skill type and posttest effect size. Models were fit using the "metafor" package in R, with a study-level random effect, random slopes for posttest effect size, weighting, and robust standard errors. Heterogeneity statistics are presented for the null model and the three subsequent models.

Table S17

Regression Model with Treatment and Control Group Growth

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Intercept | 0.15 (0.11) | -0.01 (0.10) | 0.11 (0.15) | -0.14 (0.12) |
| Posttest | 0.48 (0.18) * | 0.40 (0.14) * | 0.57 (0.24) | 0.64 (0.16) ** |
| Control growth | -0.07 (0.10) | | -0.03 (0.13) | |
| Treatment Growth | | 0.10 (0.09) | | 0.22 (0.11) |
| Control growth * Posttest | | | -0.06 (0.09) | |
| Treatment Growth * Posttest | | | | -0.18 (0.08) |
| *N* (aligned groups/studies/treatments) | 63/15/21 | 63/15/21 | 63/15/21 | 63/15/21 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

*Note*: The unit of analysis is aligned groups of posttests and impacts at 6-to-12 months after the posttest wherein the construct, measure, subscale, and reporter were the same. Growth was generated as the follow-up minus the posttest score divided by the control group's posttest standard deviation for both control and treatment groups. Aligned groups were included if (1) means and standard deviations were reported in raw (not standardized) units, (2) measures did not change units between posttest and follow-up, (3) means were not determined by averaging across groups or papers, and (4) neither control nor treatment group growth was negative nor greater than 4 standard deviations. In total, 112 aligned groups were dropped for these reasons. Coefficients can be interpreted in effect size units (i.e., the underlying data are in the effect size units reported by studies). Models were executed using the "metafor" package in R and included a study-level random effect, random slopes for posttest effect size, weighting, and robust standard errors (shown in parentheses).

Table S18

Assessing Bias: Average Posttest Effect Size Conditional on Non-Missing Data at Follow-up Wave

| | **Constrained or Unconstrained** | | **Constrained** | | **Unconstrained** | |
|---|---|---|---|---|---|---|
| | **Avg. Posttest Contingent on Follow-Up (SE)** | ***n*** | **Avg. Posttest Contingent on Follow-Up (SE)** | ***n*** | **Avg. Posttest Contingent on Follow-Up (SE)** | *n* |
| **Panel A: Weighted Average Effect Sizes** | | | | | | |
| Posttest | 0.43 (0.06) *** | 184 | 0.44 (0.07) *** | 131 | 0.34 (0.06) *** | 53 |
| 6 months to 1 year | 0.44 (0.07) *** | 175 | 0.43 (0.07) *** | 127 | 0.36 (0.08) *** | 48 |
| >1 year, up to 2 years | 0.54 (0.13) ** | 19 | 0.39 (0.11) * | 8 | 0.50 (0.18) | 11 |
| > 2 years | 0.48 (0.1) ** | 28 | 0.38 (0.10) * | 15 | 0.50 (0.13) ** | 13 |
| **Panel B: Unweighted Average Effect Sizes** | | | | | | |
| Posttest | 0.42 (0.06) *** | 184 | 0.43 (0.08) *** | 131 | 0.38 (0.09) ** | 53 |
| 6 months to 1 year | 0.42 (0.07) *** | 175 | 0.43 (0.09) *** | 127 | 0.39 (0.1) ** | 48 |
| > 1 year, up to 2 years | 0.53 (0.17) * | 19 | 0.45 (0.08) * | 8 | 0.59 (0.32) | 11 |
| > 2 years | 0.51 (0.13) * | 28 | 0.42 (0.12) | 15 | 0.61 (0.24) | 13 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

*Note*: "ES" = effect sizes. The average posttest effect sizes for aligned groups with non-missing follow-up effects at each binned wave are presented. Observing the change in posttest effect for each set of outcome estimates present at a given wave indicates the extent to which follow-up impacts could be biased by selection on large posttest effects. In Panel A, average effects included a study-level random effect, random slopes for posttest effect size, weighting, and robust standard errors. In Panel B, average effects were determined using a fixed-effects meta-analytic model with no random or econometric fixed effect for study ID. All estimates were executed in R using the "metafor" package.

Table S19

Assessing Bias: PEESE Test

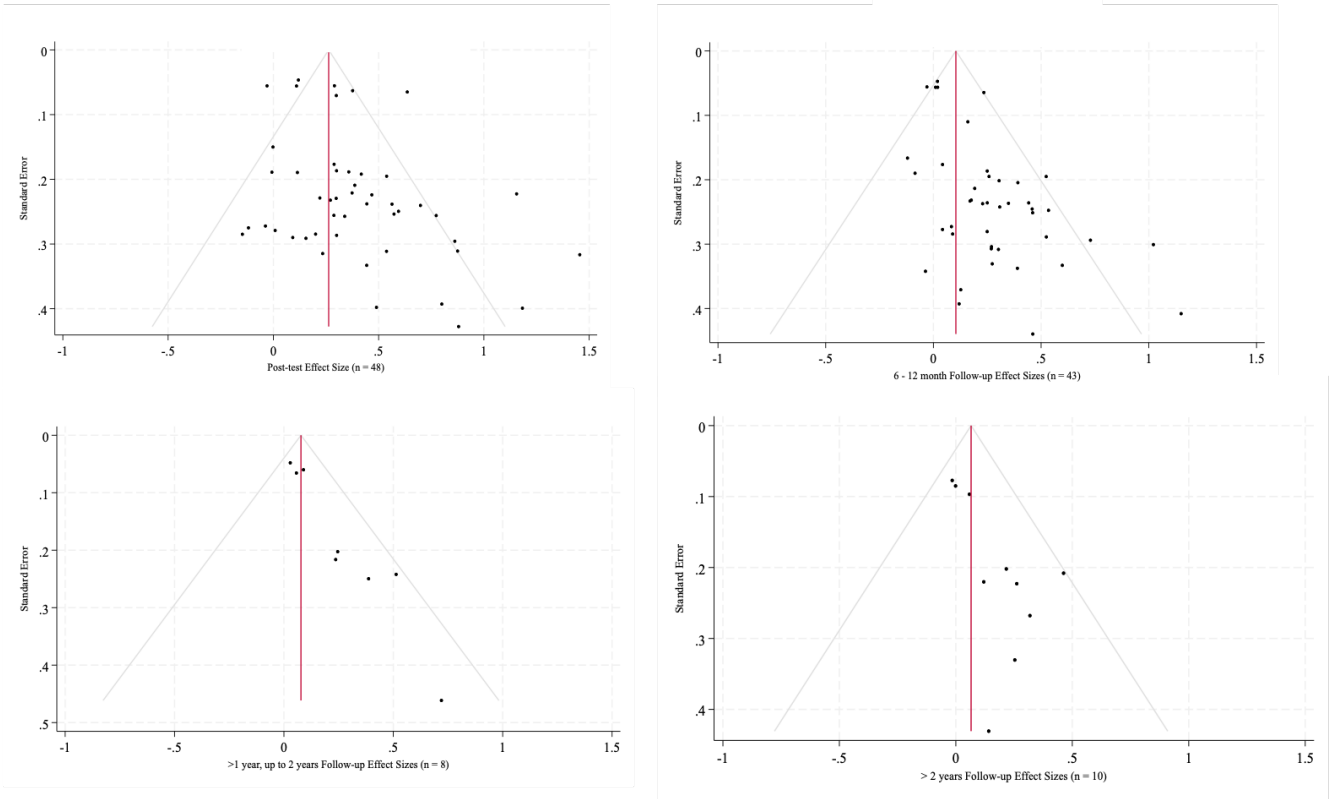
| | PEESE Test |
|---|---|
| Intercept | -0.07 (0.05) |
| Posttest | 0.45 (0.05) ** |
| Unconstrained | 0.03 (0.03) |
| Posttest x Unconstrained | -0.21 (0.08) * |
| Standard Error | 0.72 (0.26) * |
| *N* (study/tx contrasts/outcomes) | 175/43/31 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: The above table shows the results of including the estimated standard error when predicting the 6-to-12-month follow-up impact using the posttest impact. Models were fit using the "metafor" package in R, with a study-level random effect, random slopes for posttest effect size, weighting, and robust standard errors. Standard errors are presented in parentheses, and coefficients can be interpreted in effect size units (i.e., the underlying data are in the effect size units reported by studies).

Figure S1

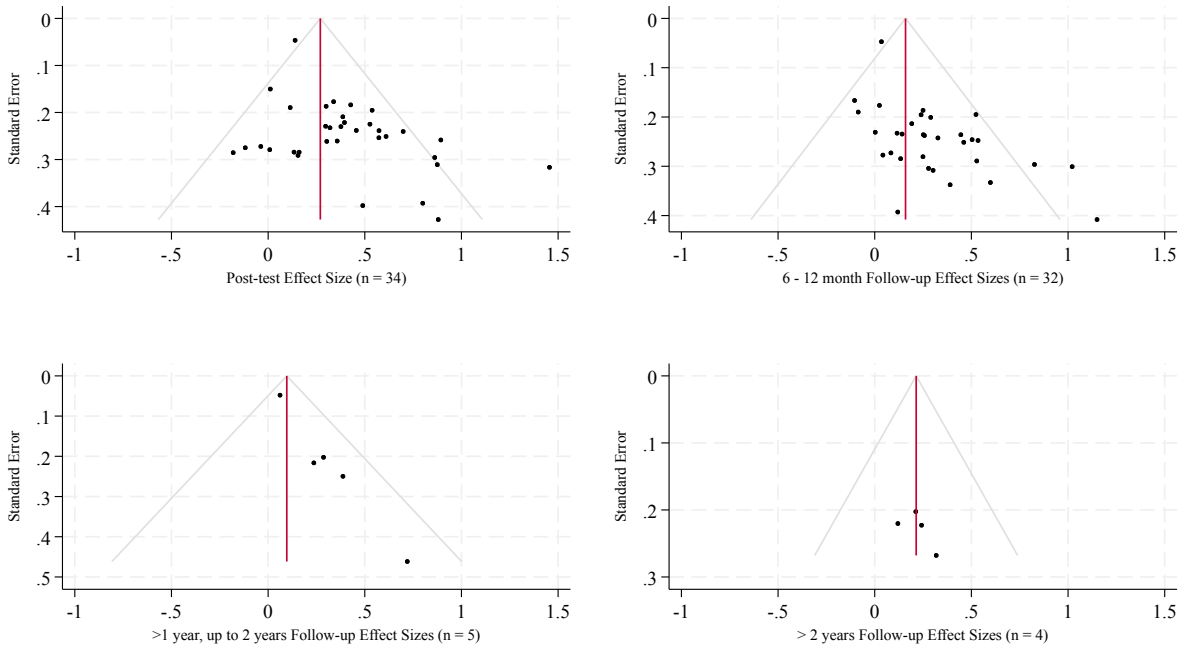
Funnel Plots of Study Effect Sizes and Standard Errors




*Note*: In the above figure, the average effect size and average standard error from each study are plotted. Each point represents one study in the analytic sample. The sample was restricted to studies that contributed at least one constrained or unconstrained aligned group.

Figure S2

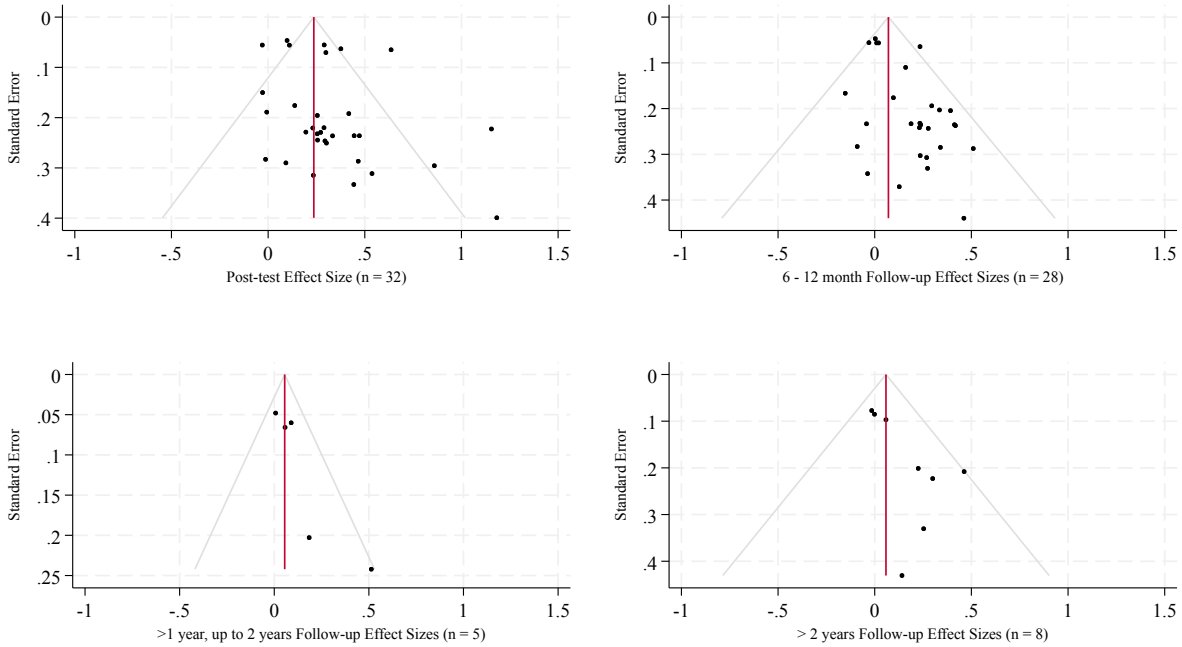
Funnel Plots of Study Effect Sizes and Standard Errors for Constrained Skills




*Note*: In the above figure, the average effect size and average standard error from each study are plotted. Each point represents one study in the analytic sample. The sample was restricted to studies that contributed at least one constrained aligned group.

Figure S3

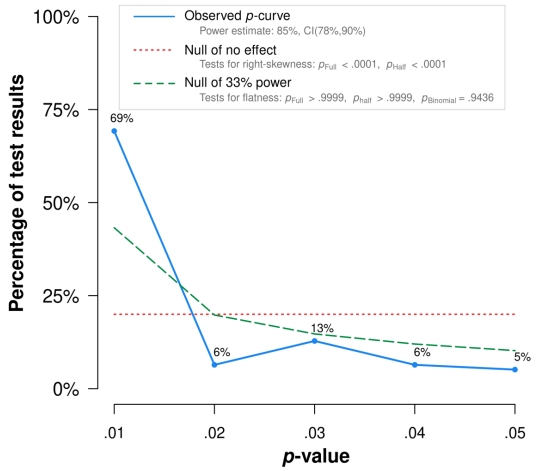
Funnel Plots of Study Effect Sizes and Standard Errors for Unconstrained Skills



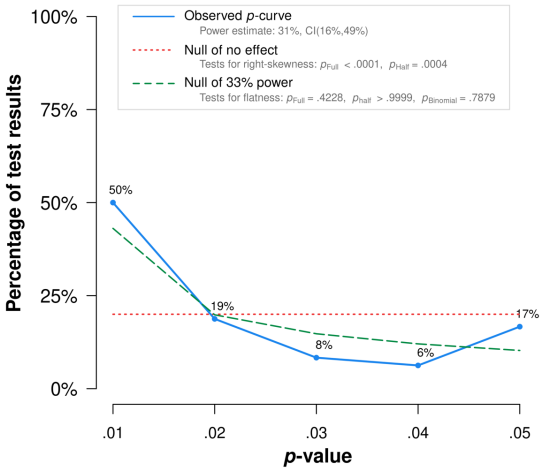

*Note*: In the above figure, the average effect size and average standard error from each study are plotted. Each point represents one study in the analytic sample. The sample was restricted to studies that contributed at least one unconstrained aligned group.

Figure S4

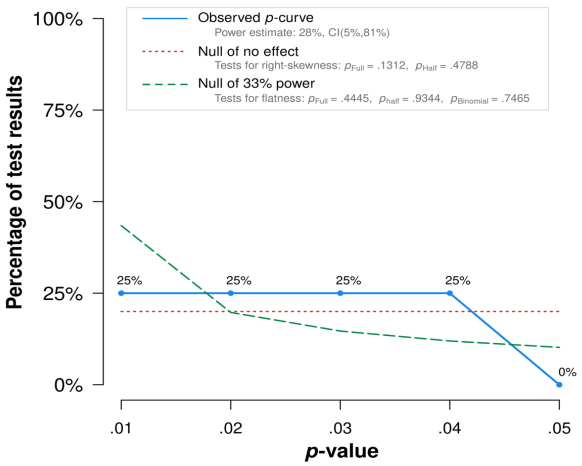
Percentage of *p*-values below 0.05 at Each Assessment Wave



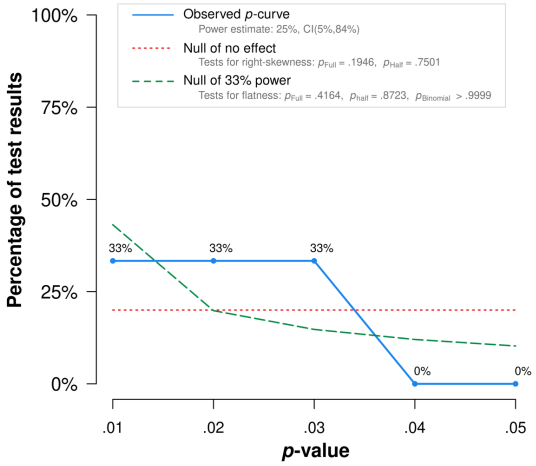

Posttest ($n$ = 184)




6 – 12 month follow-up ($n$ = 175)




1 - 2 year follow-up ($n$ = 19)

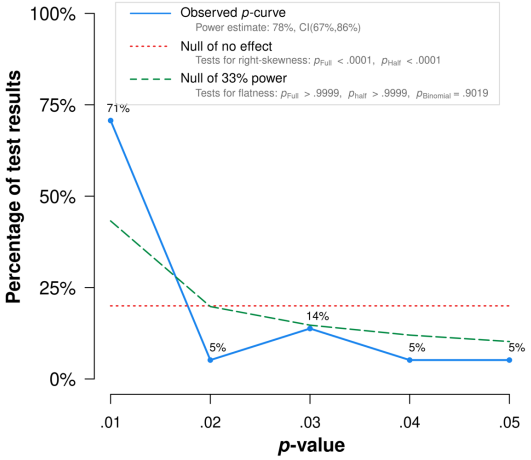



> than 2 years follow-up ($n$ = 28)

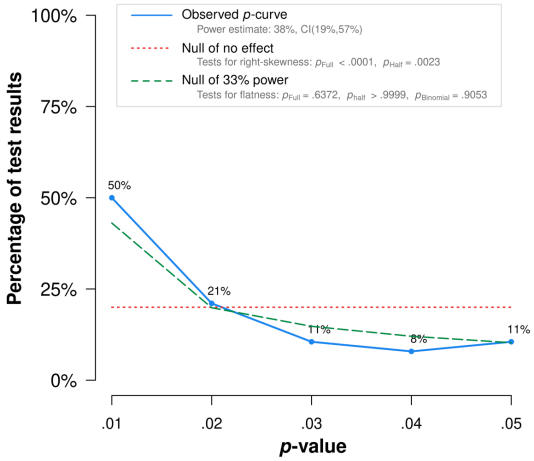
*Note*: The above p-curves only show aligned groups coded as either constrained or unconstrained. P-curves were made using https://p-curve.com/app4/.

Figure S5

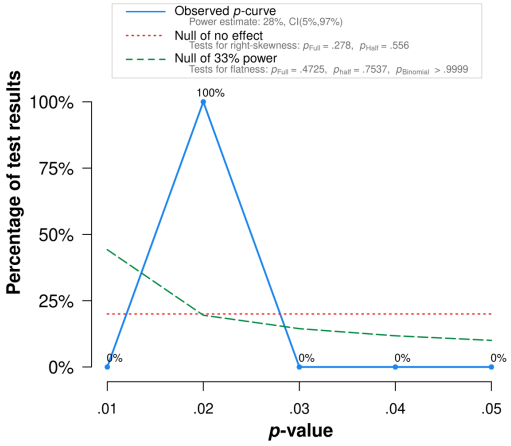
Percentage of *p*-values below 0.05 at Each Assessment Wave Among Aligned Groups Coded as Constrained



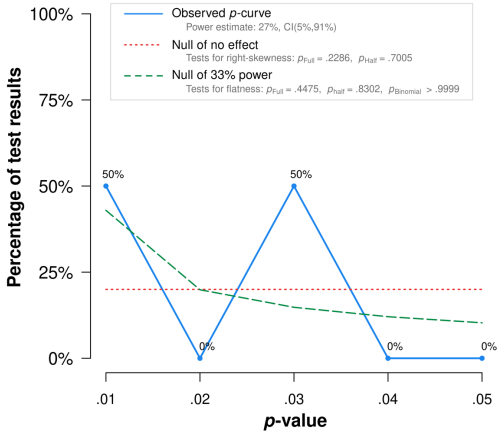

Posttest ($n$ = 131)




6 - 12 month follow-up ($n$ = 127)




1 – 2 years ($n$ = 8)

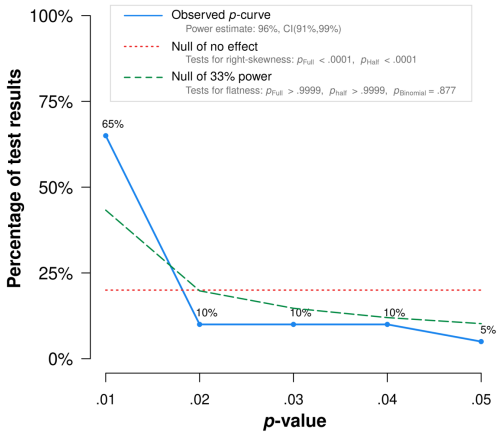



> than 2 years follow-up ($n$ = 15)

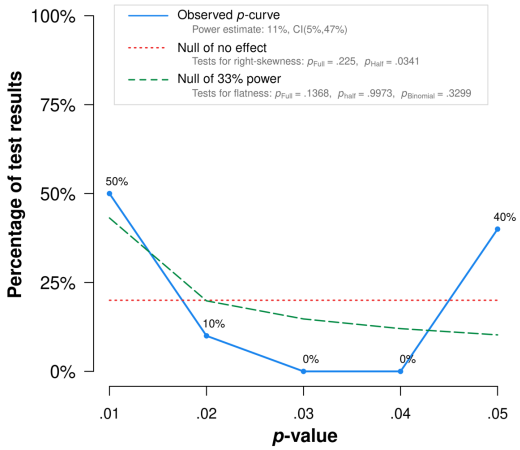
*Note*: The above p-curves only show aligned groups coded as constrained. P-curves were made using https://p-curve.com/app4/.

Figure S6

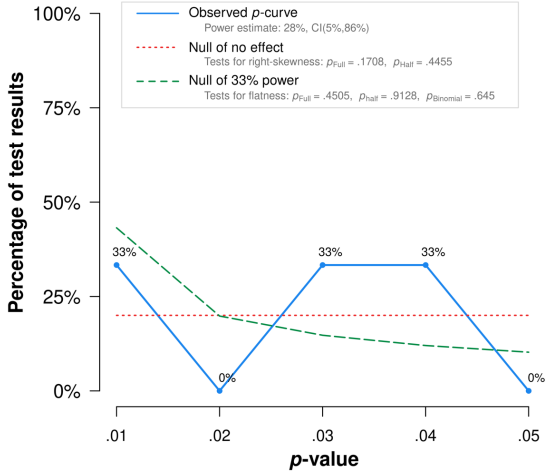
Percentage of *p*-values below 0.05 at Each Assessment Wave Among Aligned Groups Coded as Unconstrained



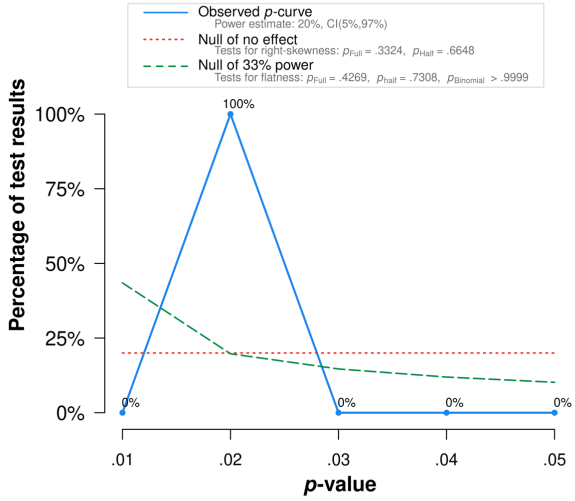

Posttest ($n$ = 53)




6 - 12 month follow-up ($n$ = 48)




1 - 2 year follow-up ($n$ = 11)




> than 2 years follow-up ($n$ = 13)

*Note*: The above p-curves only show aligned groups coded as unconstrained. P-curves were made using https://p-curve.com/app4/.