Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1367

Does Fadeout Preclude Long-Run Adult Effects? A Systematic Review and Meta-Analysis of Educational Intervention Randomized Controlled Trials

Emma R. Hart
Boston College

Drew H. Bailey
University of California, Irvine

Casey Moran
Teachers College, Columbia University

Susan Kruglinski
University of Texas at Austin

Tyler W. Watts
Teachers College, Columbia University

A perplexing pattern has been observed in the intervention literature: although impacts on children's skills often fade, some interventions still produce long-run impacts on adult outcomes. Existing developmental theory does not provide a straightforward explanation. This "fadeoutemergence paradox" spotlights our limited understanding of how early skill gains shape long-run trajectories. In part one of this paper, we reviewed the literature on childhood interventions and adult outcomes, including theory, areas of complexity, and sources of confusion. In part two, we presented a random-effects meta-analysis of randomized controlled trials that measured initial impacts on child skills and long-run impacts on adult outcomes. We identified 29 interventions reporting 179 posttest effects and 497 adult effects for 26,076 participants. On average, impacts on child skills faded in the two years after interventions ended, but programs nonetheless generated small effects on adult outcomes (.04 SD, 95% PI [.00, .09]). Strikingly, adult impacts were similar in magnitude to the faded medium-term effects, inconsistent with common conceptions of the "fadeout-emergence paradox." We observed some, albeit weak, correspondence between initial impacts and long-run impacts. Our work illustrates gaps in our understanding of how children's skills shape adult outcomes and suggests that developmental theories should accommodate fadeout followed by long-run impacts.

VERSION: September 2026

Suggested citation: Hart, Emma R., Drew H. Bailey, Casey Moran, Susan Kruglinski, and Tyler W. Watts. (2026). Does Fadeout Preclude Long-Run Adult Effects? A Systematic Review and Meta-Analysis of Educational Intervention Randomized Controlled Trials. (EdWorkingPaper: 25-1367). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/dtk5-ds56

Draft version 2, 8/3/26.
This paper has not been peer reviewed.

Does Fadeout Preclude Long-Run Adult Effects? A Systematic Review and Meta-analysis of Educational Intervention Randomized Controlled Trials

Emma R. Hart[1], Drew H. Bailey[2], Casey Moran[3], Susan Kruglinski[4], Tyler W. Watts[3]*

[1] Boston College; [2] University of California, Irvine; [3] Teachers College, Columbia University; [4] University of Texas at Austin

Corresponding Author:
Tyler W. Watts
462 Grace Dodge Hall
525 W 120th
New York, NY 10027
(212) 678-3095

Acknowledgments

Time spent on this project was supported by the National Institute of Child Health and Human Development (1R01HD095930-01A1 to TW) and the National Science Foundation (DGE-2036197 to EH). Thank you to Eric Dearing, Jens Dietrichson, Greg Duncan, James Pustejovsky, Beth Tipton, Mikkel Helding Vembye, and Michael Weiss for their helpful feedback on earlier versions of this manuscript. We would like to thank the following research assistants whose efforts made this work possible (in alphabetical order): Abby Lee, Ana Duenas Crespo, Camila Torres Rivera, Fatmanur Ozay, Katarina Djordjevic, Siqi Liang, Yiming Wang, Yolanda Xu, and Victoria Chang.

Author Note

Prior to publishing, the codebooks, analytic syntax, and data necessary to replicate these findings will be publicly posted. Contact the authors for these materials in the meantime. This study was not pre-registered. An earlier version of the paper is posted at Annenberg Ed Working Papers (https://edworkingpapers.com/ai25-1367). Earlier versions of this work were also presented at the SRCD, SREE, and Stanford Educational Opportunity conferences.

Abstract

A perplexing pattern has been observed in the intervention literature: although impacts on children's skills often fade, some interventions still produce long-run impacts on adult outcomes. Existing developmental theory does not provide a straightforward explanation. This "fadeout-emergence paradox" spotlights our limited understanding of how early skill gains shape long-run trajectories. In part one of this paper, we reviewed the literature on childhood interventions and adult outcomes, including theory, areas of complexity, and sources of confusion. In part two, we presented a random-effects meta-analysis of randomized controlled trials that measured initial impacts on child skills and long-run impacts on adult outcomes. We identified 29 interventions reporting 179 posttest effects and 497 adult effects for 26,076 participants. On average, impacts on child skills faded in the two years after interventions ended, but programs nonetheless generated small effects on adult outcomes (.04 SD, 95% PI [.00, .09]). Strikingly, adult impacts were similar in magnitude to the faded medium-term effects, inconsistent with common conceptions of the "fadeout-emergence paradox." We observed some, albeit weak, correspondence between initial impacts and long-run impacts. Our work illustrates gaps in our understanding of how children's skills shape adult outcomes and suggests that developmental theories should accommodate fadeout followed by long-run impacts.

Keywords: educational RCTs, long-run effects, fadeout, meta-analysis, developmental theory

Public Significance Statement

We found that interventions improve children's skills at program end, that these improvements fade considerably over time, but that programs nonetheless generate small, positive impacts on adult functioning. The extent to which a program initially improved children's skills showed weak correspondence with the extent to which it improved adult outcomes.

Does Fadeout Preclude Long-Run Adult Effects? A Systematic Review and Meta-analysis of Educational Intervention Randomized Controlled Trials

"Well, everything dies, baby, that's a fact, but maybe everything that dies someday comes back"
– Bruce Springsteen

In 2023, Gray-Lobe and colleagues published an evaluation of the long-run effects of Boston's preschool program. The findings were theoretically perplexing: children randomized to receive a year of preschool had no apparent advantages in elementary school achievement or behavior but had higher rates of high school graduation and college attendance years later. The study suggested the possibility of *emergence*, that education could generate effects on life course outcomes despite apparent null effects on measures of child skills. Gray-Lobe's is not the only educational intervention to find long-run impacts despite no, or fading, impacts on child skills (e.g., Deming, 2009; Chetty et al., 2011). How could emergence be possible?

Emergent impacts of educational interventions can be understood as a type of 'sleeper effect,' which has a long history in psychology. The concept was formally introduced about 75 years ago when Hovland et al. (1949) found that although individuals initially discounted information provided by untrustworthy sources, some eventually came to accept it, much to the researchers' surprise. Perhaps similarly, in human development, "U-shaped" patterns of development, where "children do worse before they do better" (Gershkoff-Stowe & Thelen, 2004, p. 11), have continually perplexed researchers (Siegler, 2004). Some psychologists have theorized that continuous underlying processes may manifest as seemingly discontinuous behavioral expressions across development (e.g., "heterotypic continuity"; Kagan, 1971). Even Freud (1922) argued that childhood experiences can lead to neuroses that only emerge in adulthood. Indeed, researchers have long debated whether these apparent sleeper effects reflect real and robust phenomena or non-replicable patterns affected by methodological issues (e.g., Clarke & Clarke, 1981; Cook et al., 1979; Klahr, 1982).

In the current paper, we attempted to improve the field's understanding of whether the findings in Gray-Lobe et al. (2023) are representative of a more general and widespread pattern in the educational intervention literature, and whether we can predict when interventions will find long-run impacts based on initial impacts on child skills. In part 1 of the paper, we review the literature. We begin by detailing developmental theories describing how child skills shape adult outcomes. We then discuss how these theories have been complicated by intervention fadeout and theoretical explanations that have been offered as a result. Finally, we discuss why it remains difficult to draw clear conclusions about long-run effects from the current literature.

In part 2, we report the results from a meta-analysis of the short- (i.e., end of the intervention), medium- (i.e., six months to two years later), and long-run adult effects of educational interventions evaluated using randomized designs (i.e., RCTs). We included programs that targeted children's skills prior to high school and thus arrived at a diverse sample of 29 interventions, including evaluations of nurse home visiting, early education, reading remediation, and substance use prevention. We refer to these programs as "educational" because they attempted to improve children's skills. Yet, the programs used theoretical models derived from various subfields of psychology. Drawing on this sample, we first set out to identify the basic trajectory of longitudinal intervention impacts and whether the pattern of fadeout followed by emergence is broadly observed. We then tested whether short-term impacts on child skills forecasted long-term impacts on adult outcomes.

Part 1: Review

The Case for Foundational Childhood Skills

The idea that prior skill acquisition shapes subsequent skill advancements is a central assumption of developmental theory (Cunha & Heckman, 2007; Masten & Cicchetti, 2010). In the realm of cognitive development, this idea is particularly intuitive. For example, basic decoding skills are thought to help children to learn more from reading, thereby supporting more advanced literacy skills (Share, 1995). Prior skills may not only shape subsequent within-domain functioning but also shape functioning in other domains, a process described as *transfer*. For example, behavioral regulation may additionally impact a child's ability to make literacy progress (Blair & Raver, 2015). Across longer timescales, it is expected that earlier skills can alter trajectories in ways that are commensurate with disparate adult outcomes.

Cunha and Heckman's *Technology of Skill Formation* (2007) and Masten and Cicchetti's *Developmental Cascades* (2010) are two influential expressions of these ideas. Both theories establish the expectation that differences in early skills will persist and ultimately shape adult functioning. Captured in the maxim that "skills beget skills," economists Cunha and Heckman argued that stronger earlier skills lead to stronger later skills through mutually reinforcing self- and cross-productivities within and across domains. They also argued for dynamic complementarities as a critical mechanism driving skill growth, whereby stronger skills enable better use of subsequent learning environments. Developmental psychologists Masten and Cicchetti similarly argued that development occurs through cascades, in which functioning in one area shapes functioning in the same and other areas through chain-like reactions. The result of such processes can be likened to the "Matthew Effect," where 'the rich get richer' (Perc, 2014). Although the basic premises of these theories are generally accepted, details remain debated, such as the network-level processes by which growth occurs (e.g., Mak & Twitchell, 2020; van der Maas, 2024) and the extent to which transfer happens (Barnett & Ceci, 2002).

Still, the basic idea that early skill gains support later skill advancement seems non-controversial and has found broad support in correlational studies. Indeed, empirical work regularly demonstrates that individual differences in skills are stable across time (e.g., Breit et al., 2024): children with more advanced skills early in life tend to have more advanced skills later in life. Other influential correlational work has also established convincing links between child skills and adult competencies (e.g., Jones et al., 2015). Together, correlational findings have often been taken to suggest that intervening to change children's skills will result in sustained benefits and improved adult outcomes. For example, Moffitt et al. (2011) observed that a measure of childhood self-control was associated with adult health, socioeconomic, and public safety outcomes and, thus, concluded that interventions targeting self-control might "reduce a panoply of societal costs, save tax payers money, and promote prosperity" (p. 2693).

As an example of an application of this rationale, consider the Good Behavior Game, a classroom behavioral management intervention for first graders (Poduska et al., 2008). One could argue that behavioral problems in children as young as first grade do not yet warrant intervention. Most children will grow out of it, and those who do not can receive intervention at later ages. Yet, the skill intervention perspective informed by developmental theories suggests that early intervention is better, because early behavioral skill deficits will compound with time. Indeed, the Good Behavior Game was influenced by correlational work showing that measures of aggressive and shy behaviors in early childhood predict later psychiatric issues and antisocial behaviors (e.g., Kellam et al., 1980). The program itself targeted teachers' classroom practices with the ultimate goal of influencing the development of childhood aggression. The investigators anticipated that by changing behavioral issues at the time when they become predictive of later outcomes (i.e., early childhood), they could lock in improved long-term trajectories.

These types of optimistic predictions for interventions targeting children's skills are highly prevalent in the literature. In Figure 1, we have attempted to depict plausible long-term effect predictions by stylizing several prominent patterns of longitudinal effects that could unfold following a skill-focused childhood intervention. Necessarily, Figure 1 is an oversimplification, but the archetypes we present are useful for understanding the predictions in the literature.

The most optimistic patterns in Figure 1—that intervention-driven boosts to child skills will persist ("Full Persistence") or grow ("Growing Effects")—provide the strongest support for investment in interventions. Such patterns may be possible under strong dynamic complementarities and cross-productivities. Figure 1 also depicts a more moderate version of persistence, "Considerable Persistence," supported by correlational studies showing strong, but imperfect, stability in skills over time (see Bailey et al., 2018). In all three of these cases, one would predict that intervention-driven boosts to child skills would be detectable for years in the future, with initial benefits leading to long-term advantages for child skills and adult functioning.

The Fadeout Problem

In recent years, research has challenged optimistic expectations of long-term skill advantages following interventions (Bailey et al., 2020). Rigorous evaluations of skill-focused interventions have often found that although children who received an intervention tend to have stronger skills than those who did not at the program's end, benefits often fade in the months and years that follow (Bailey et al., 2020; Hart et al., 2024; Protzko, 2015). Importantly, this apparent fadeout happens at faster rates than predicted by correlational research (Bailey et al., 2018). The prominence of fadeout in the intervention literature raises concerns about our ability to generate persistent intervention effects (Bailey et al., 2024) and the validity of using correlational estimates to forecast intervention impacts (Brick & Bailey, 2020).

Theory and correlational findings would lead us to expect that if children who did and did not receive an intervention are no different in the years after the program ends, then we should also find no differences in adulthood (e.g., "Full Fadeout" in Figure 1). However, Gray-Lobe's (2023) evaluation of Boston pre-k found emergent long-run impacts *despite* null intermediary effects on child skills. Indeed, other rigorous evaluations have similarly reported adult impacts despite no initial effects or medium-term fadeout ("Sleeper Effects" or "Fadeout-Emergence" in Figure 1). For example, Deming's (2009) quasi-experimental analysis of Head Start found fadeout on cognitive skills during childhood, followed by positive effects on high school graduation, health, and the likelihood of pursuing higher education or work. Chetty et al.'s (2011) analysis of classroom quality also found long-run impacts on earnings and college attendance despite fadeout on achievement. Relatedly, Hitt et al. (2018) reported that the short-run achievement effects of school choice programs were apparently unrelated to their long-term effects on educational attainment.

Theoretical Accounts of Fadeout and Emergence

Fadeout of initial effects followed by adult impacts, which we refer to as the "fadeout-emergence paradox," is not a pattern that developmental theorists or interventionists generally anticipate. Although the pattern is not strictly "rare" in the education intervention literature, it has been continually met with some mix of excitement, skepticism, and bewilderment by observers.[1] The fadeout-emergence paradox raises questions about how interventions generate long-run effects if not through persistent impacts on targeted child skills.

Can existing developmental theory reconcile the pattern of fadeout and emergence? At first blush, the theories reviewed above (Cunha & Heckman, 2007; Masten & Cicchetti, 2010) do not appear to explicitly predict fadeout, nor do they provide an easy way of understanding how fading skill impacts could be followed by adult impacts. Indeed, these theories are usually used

---

[1] Table S1 provides a list of relevant quotes.

by scholars to make vague predictions that their intervention will generate persistent effects (see Table S2). However, like many psychological theories, theories of skill building are flexible, leaving room for most empirical results to be interpreted as broadly consistent with theory.

Yet, the current explanations on offer for fadeout and emergence still seem to hinge on the persistence of some kind of skill impact. The most prominent explanation is that programs must have generated persistent impacts on unmeasured behavioral skills. Drawing on developmental theory and correlational evidence (e.g., Duckworth et al., 2018; Jones et al., 2015; Moffitt et al., 2011), some have claimed that interventions are likely to generate adult impacts due to overlooked and persistent effects on social-emotional or "non-cognitive" skills (Chetty et al., 2011; Deming, 2009; Heckman et al., 2013; Gray-Lobe et al., 2023). Gray-Lobe et al. (2023) seem to prefer this explanation when interpreting their emergent impact results (see Table S1).

However, a recent meta-analysis of 86 educational RCTs casts some doubt on this explanation. Hart et al. (2024) found that posttest impacts on social-emotional skills faded at a similar rate (~50%) as impacts on cognitive skills at 6-month to 1-year follow-up. Although this does not rule out the possibility that unmeasured social-emotional skill persistence (or partial persistence) could produce some of the adult impacts observed in the literature, the results suggest that, at least in the short term, persistent social-emotional impacts might be as rare as persistent cognitive impacts. Moreover, other studies have found that measured social-emotional skills generally account for less than 50% of adult program impacts (Elango et al., 2016; Pages et al., 2023). More complex mechanisms must be at work.

When searching the theoretical literature for potential explanations, one finds many possibilities, but few theories lead to predictions that can be rigorously tested. Indeed, many theories articulate the complexity of development where skills and life outcomes are shaped through interactions between the individual and their context (e.g., Bronfenbrenner, 1992; Mak & Twitchell, 2020; Thelen, 2005; Werner, 1957). Because these theories describe development as a highly complex process, it does not seem impossible that programs could generate long-run effects despite fadeout. Like the theories reviewed above (Cunha & Heckman, 2007; Masten & Cicchetti, 2010), these theories are flexible enough that they could be consistent with nearly any pattern of effect and, indeed, we cannot think of any theory that explicitly argues against fadeout as a possibility. However, we are also unaware of any theory that predicts fadeout as a normative pattern and, in practice, researchers almost never predict fading intervention effects. Whereas theories have been successfully applied to create programs that initially benefit children in predictable ways, fadeout and emergence raise concerns about whether research can inform programs that yield dependable long-run benefits, or whether the longitudinal patterns of program effects are too theoretically underdetermined (Cronbach, 1969).

As a step toward formalizing a developmental model that explicitly reconciles fadeout in the medium term with longer-term adult impacts, Bailey et al. (2024) recently proposed the Large Interconnected Network Theory (LINT). LINT depicts fadeout as a regularity and articulates how interventions can produce longer-term impacts despite fadeout. LINT argues that although impacts on any one targeted skill may diminish quickly, intervention impacts on targeted child skills could influence other child skills, as well as their subsequent contexts, eventually leading to the formation of a small network-level impact that stabilizes through ongoing interactions over time. As depicted in Figure 1, interventions could generate impacts on adult outcomes that are similar in magnitude to the intervention's medium-term effect on child skills ("Network-level effects (LINT)" in Figure 1). A simple version of LINT in which skills affect each other symmetrically in a linear system over many iterations suggests the usefulness of a skill-type null hypothesis: because impacts on skills converge to some non-zero level over time[2], it may be difficult to pinpoint which specific intervention-targeted skills will most heavily shape adult outcomes. This version of LINT could be described as a specific parameterization of general skill-building models. LINT provides an explicit theoretical accounting for fadeout and emergence through the kinds of skill-to-skill transfer processes and complex interactions with context that many existing theories postulate (e.g., Bronfenbrenner, 1992; Cunha & Heckman, 2007; Mak & Twitchell, 2020; Masten & Cicchetti, 2010; Thelen, 2005; Werner, 1957).

The Current Literature on Long-term Effects

In addition to issues with theory articulation, limitations of the intervention evaluation literature have also set serious roadblocks for forming a robust intervention science. It is challenging to test causal skill-building processes with a single intervention study because point-in-time estimates cannot reveal the mechanisms driving long-run effects (Rohrer & Lucas, 2023). Programs are often "fat-handed" in their targets, plausibly acting on many mechanisms (Eronen, 2020). Moreover, few studies conduct follow-up, and those that do likely include non-representative samples that produced large and promising initial effects (Watts et al., 2019).

Variation in how researchers report follow-up impacts has also made it difficult to decipher and synthesize results. Researchers may selectively report outcomes, change which skills they measure over time based on earlier results, report impacts on different subgroups across papers, and draw theoretical conclusions based on a fraction of their total findings (Pigott et al., 2013; von Hippel & Schuetze, 2025; Wagenmakers et al., 2012). If follow-up effects are

[2] In the very long term, impacts in a stable linear system (i.e., one where variances are constant across time) will eventually converge to 0. However, it is possible for impacts to persist at a substantially higher level across the lifespan, depending on the exact parameter values in the system. Simulations of effects across time in a complex linear system are shown in Bailey et al. (2024); the code can be used to probe the predictions of models with different parameter values.

highly selected by researchers, the same study might be consistent with several of the patterns in Figure 1. Take, for example, the Perry Preschool Program, an intervention that has exemplified the fadeout-emergence paradox and has been highly influential in the field. Anderson (2008) critiqued the inferences drawn from different analyses of the data, flagging inconsistencies in the modeling approach and the use of gender as a moderator. For example, Anderson's re-analysis found weak support for impacts on crime among boys, whereas past work highlighted robust impacts and argued that such effects drove Perry's public returns (Schweinhart et al., 2005).

Depending on what outcomes researchers track over time and how they synthesize evidence across outcomes, they may come to completely different "stories" about an intervention's mechanisms. For example, Perry's posttest impacts on IQ—the primary outcome of focus when the intervention was developed—famously faded to approximately 0 by third grade (Schweinhart & Weikart, 1981). Pairing this apparent fadeout with the long-term impacts on crime mentioned above, for instance, a researcher could easily find support for the fadeout-emergence paradox trajectory ("Fadeout-Emergence" in Figure 1), whereas aggregating effects across the full range of outcomes measured at different waves could lead to conclusions that the results resemble something more like the "Network-level (LINT)" trajectory in Figure 1. Indeed, although impacts on IQ faded, effects on some social-emotional skills were more persistent (Heckman et al., 2013; Schweinhart et al., 2005), and impacts on a summary index of adult outcomes were smaller than effects on criminal activity for males (Anderson, 2008).

Thus, despite the widespread interest and clear importance of intervention evaluations for psychological theory and public policy, this area of research is limited by issues in both theory articulation and study design. When researchers delineate specific mechanisms they expect to drive their intervention's long-run effects, these theories are generally forgiving, fit a variety of observed results, and rarely yield clear falsification tests (Eronen & Bringmann, 2021; Oberauer & Lewandowsky, 2019). The lack of theoretical clarity is then further reinforced by the design of long-run intervention studies, where follow-up is pursued only in selected cases, many outcomes are collected when follow-up is undertaken, and researchers face a plethora of analytic choices that allow for ever-expanding degrees of freedom. It is thus challenging to decipher what lessons to draw from any one study's results, and to know if the field is making progress toward a deeper understanding of human development or cycling through *post hoc* explanations of noisy results that are unlikely to amount to a robust and falsifiable science. A different approach is warranted.

Part 2: Meta-analysis

In the current study, we endeavored to provide a straightforward accounting of the educational intervention literature using meta-analysis. We attempted to gather all educational intervention RCTs conducted from infancy to middle school, with impacts evaluated at the end of the intervention and in adulthood, to improve our understanding of basic patterns, namely: Is the fadeout-emergence pattern broadly observed across different types of interventions? Do short- and medium-term effects provide meaningful information about long-run impacts?

To be clear, we did not set out to test highly specific theories or adjudicate among alternative mechanisms. We also did not hope to generate a definitive meta-analytic estimate of the long-run effect of educational interventions, nor to identify all interventions that have shown sleeper effects. Rather, we sought to broadly inform current developmental theories by observing what is generally true across a diverse set of interventions. From the outset, we knew that the universe of studies that measured adult follow-up was small and heterogeneous (Watts et al., 2019). Considering the remarkable influence that some small RCTs with long-run follow-up have had on research and policy (e.g., Perry Preschool), our goal was to provide a comprehensive and accurate synthesis of the state of this important, albeit small, literature.

Method

Data

We created a new meta-analytic sample for this study, following many of the methods in Authors (XXXX). We provide a brief overview here and more details in the supplement.

Inclusion Determinations. Figure 2 outlines the flow of papers reporting results for unique studies through our inclusion/exclusion decision-making process. In February 2024, the first author conducted database searches across PsycINFO, ERIC, and EconLit, yielding 8,167 papers. Covidence identified 7,201 non-duplicate papers. At least one research assistant and the first author then independently screened abstracts for adherence to several criteria: (1) reported long-term follow-up impacts on adult outcomes, (2) used RCT or lottery design, (3) aimed to improve child or adolescent cognitive and/or social-emotional development, and (4) took place in the United States (see supplement for discussion). 155 papers passed abstract screening.

Next, the first author reviewed each study's assessment schedule and intervention timing to identify final inclusion criteria. Here, we determined two additional criteria: interventions had to (1) end prior to students' entry to high school and (2) contain follow-up assessments after age 18 or grade 12. At least one research assistant and the first author independently conducted a full-text review of the 155 papers. 80 papers reporting results from 28 unique studies passed (there were many papers per study). We compared the included studies to those included in a recent meta-analysis of the long-run effects of educational interventions by Hart et al. (2024) that relied on a different search protocol. We identified four additional studies that met the inclusion criteria and included these four additional studies at this stage (see supplement for details).

We then gathered all papers reporting intervention effects for these 32 studies (28+4). For each paper reporting on these 32 studies (84 papers), two research assistants independently conducted a series of "forward" and "backward" searches to identify other papers that contained posttest or follow-up impacts. The research group then worked together to make final inclusion determinations based on whether the study included posttest and follow-up data. We found that 25 studies with 29 treatment-control-group contrasts generated through random assignment (henceforth referred to as "interventions") met the inclusion criteria. Among the 882 papers we found for these 25 studies, we identified 94 unique papers that reported posttest, 6-month to 2-year follow-up, and/or adult follow-up effects, which we then double-coded (see supplement).

The sample was challenging to code due to unclear reporting within and across papers. Some violations of the inclusion criteria were identified after we started coding. In many cases, problematic study details were inconsistently reported across papers. To our knowledge, two interventions provided booster sessions in high school, three interventions dropped some participants (based on cohort, race, preschool attendance), four interventions reported posttest impacts collected prior to the end of the primary interventions, and two interventions reported posttest impacts measured with a delay after intervention end. Details on the deviations are provided in Table S3. We retained these studies in our sample, but probed intervention-specific issues through robustness checks including leave-one-out models that dropped each intervention.

Coding. All 94 papers were independently double coded by the first and third authors. Prior to double-coding each paper, coders reached 79% reliability (percentage of non-discrepant intervention- and effect-size-related codes) on a random sample of 5 papers from the 80 identified following full-text review. Prior to discrepancy checking, 89% of codes were non-discrepant across the full sample. An independent research assistant identified discrepancies, and coders reached consensus in consultation with the doctorate-level authors as needed. For each study, we coded a broad array of treatment characteristics and study features, detailed in the posted coding protocol. We coded the posttest, the 6-month to 2-year follow-up, and the adult impacts, as well as information relevant to each impact (e.g., instrument details).

Effect Size and Standard Error. We then computed standardized effect sizes and standard errors for each outcome. When effect sizes were reported in standardized units, we used the researcher-reported statistics. However, when intervention effects were reported in units other than standard deviations or were unreported, we used the available information and various approaches to estimate treatment impacts (see Figure S1). We only included main treatment effects; when subgroup impacts were reported, we computed a weighted average of impacts and standard errors using subgroup sample sizes. We rescaled all effects so that larger effects indicated preferred outcomes. When possible, we standardized treatment-control differences using control-group standard deviations, consistent with the Glass's delta approach.

Other than in the rare case that a paper provided a standard error that accompanied an effect size in standard deviation units that did not require conversion, we computed standard errors using the effect size estimate and sample sizes (Borenstein et al., 2009; p. 27):

$$SE_{ES} = \sqrt{\frac{n_{tx} + n_{cntrl}}{n_{tx} n_{cntrl}} + \frac{ES^2}{2(n_{tx} + n_{cntrl})}} \qquad (1)$$

To deflate the precision of standard errors for effect sizes from cluster-based RCTs, we adjusted the standard errors by a variance inflation factor that accounted for the average sample size per cluster and an assumed ICC of 0.10. We did not perform this adjustment if we were able to use author-reported standard errors from a model with cluster adjustments. We ran checks using alternative standard errors generated from the authors' reported data (see Figure S2).

Analysis

Descriptive Patterns of Longitudinal Intervention Impacts. First, we examined the longitudinal trajectory of intervention impacts across posttest, medium-term follow-up, and long-run follow-up assessments. Following Hart et al. (2024), medium-term follow-up assessments were categorized as 6- to 12-month follow-up and 1- to 2-year follow-up for analysis. As shown in Figure S3, on average, studies measured adult effects four times across adulthood, though as few as once and as many as eight times. Our models considered adult impacts simultaneously, regardless of timing. Sensitivity checks found that assessment timing did not impact our results.

We estimated the meta-analytic average of intervention impacts at each assessment wave (i.e., posttest, 6- to 12-month follow-up, 1- to 2-year follow-up, and adult follow-up) using R. First, we used the *metafor* package to fit random-effects meta-analytic models incorporating the sampling variances of each effect size estimate and study-level random effects (Viechtbauer, 2010). Then, we used the *clubsandwich* package to implement robust variance estimation with default CR2 small-sample-size adjustment to obtain cluster-robust standard errors, clustered at the study level (Pustejovsky, 2023; Pustejovsky & Tipton, 2018). We computed the meta-analytic average of impacts at each assessment wave using (a) all outcome-level impacts, (b) intervention-level averages of all outcome-level impacts, (c) intervention-level averages of impacts for outcomes that were consistently assessed at posttest and medium-term follow-up.

Medium-term Persistence. We next investigated whether posttest impacts persisted 6 months to 2 years following intervention end using methods detailed in Hart et al. (2024). In short, we first identified cases for which the same construct was assessed consistently using the same instrument, reporter, and subscale at posttest and at least one follow-up. In two separate models, we regressed impacts collected at 6- to 12-month follow-up and 1- to 2-year follow-up on posttest impacts. These models produced two terms: First, the slope term or "conditional persistence rate," which indicated the proportion of posttest impacts that persisted at follow-up, conditional on the magnitude of the posttest effect (i.e., slope=1 would indicate 100% persistence of posttest impacts). Second, the intercept term captured the portion of the follow-up effect unexplained by posttest effects on the same skill. We estimated the models as follows:

$$ES_{sifg} = \beta_{0s} + \beta_1 ES_{sipg} + \varepsilon_{sifg} \quad (2)$$

$$\beta_{0s} = \gamma_{00} + u_{0s} \quad (3)$$

$$u_{0s} \sim N(0, \tau^2), \quad \varepsilon_{sifg} \sim N(0, \upsilon_{sifg}) \quad (4)$$

where i represented intervention, s represented study, f indicated the follow-up wave (6- to 12-months or 1- to 2-years), p indicated posttest, and g indicated the construct that was consistently assessed. The study-level random effects, u_{0s}, were assumed to be normally distributed around 0 with between-study variance τ^2. Sampling errors were assumed to be normally distributed around 0 with variance υ_{sifg} which was the estimated sampling variance for each effect size estimate. We adjusted the estimates using robust variance estimation (as described above).

Correspondence between Posttest Impacts and Adult Follow-Up Impacts. To examine the association between posttest impacts and adult impacts, we used an adapted version of equations (2) and (3) in which we regressed average intervention-level follow-up impacts on posttest impacts. If long-run intervention impacts were caused by initial intervention effects on measured skills (and unmeasured, correlated factors), then we would expect to observe a small intercept term and a large slope term. If interventions were "black boxes," then we would expect to find a large intercept term and a small slope term (Pages et al., 2023).

Both posttest and adult impacts were measured with error. If the error were uncorrelated, we would expect a larger intercept term and an attenuated slope term. If correlated, then the slope may be an upward bound for the true correlation. We ran supplemental analyses using

Gilbert and Himmelsbach's (2026) method to probe the impact of error on our estimates under various assumptions. Importantly, the slope may also be biased by omitted variables.

Transparency and Openness Promotion

We followed PRISMA guidelines (Page et al., 2021). However, we did not conduct formal study-level risk-of-bias assessments. We relied on RCTs to improve the internal validity of our estimates, and we formally assessed confidence in our findings by examining the precision and robustness of our estimates across sensitivity and selection bias analyses. We interpreted two-sided p values less than .05 as statistically significant. The study's data, syntax, and codebook will be posted publicly prior to publication. We used R 4.5.1, and primary analyses relied on metafor 4.8.0 (Viechtbauer, 2010) and clubsandwich 0.6.1 (Pustejovsky & Tipton, 2018). We used ChatGPT to troubleshoot statistical code and to make minor edits to author-generated text (OpenAI, 2025). This study's design and analysis were not pre-registered.

Results

Details on the Sample of Educational RCTs that have Measured Child and Adult Effects

Table S3 provides details on each of the 25 studies and 29 intervention groups that comprised the sample and the supplemental text provides additional description of the studies. Interventions began as early as 1963 and as late as 2013, lasted 1 month to 8 years ($M \cong 2$ years), and, by design, targeted children as early as in utero and as late as at the end of middle school ($M = 7$ years old). Samples ranged from 33 to 10,170 children ($M = 916$). Interventions in our sample were generally broad in scope: 72% targeted parents, 93% targeted social-emotional skills, and 48% also targeted cognitive skills. Only 1 intervention targeted just cognitive skills.

Table S4 summarizes our sample of 179 posttest impacts and 497 adult impacts. On average, each intervention reported 6 posttest impacts (range = 1 to 25) and 28 adult follow-up impacts (range = 1 to 60). Adult effects were assessed at 1 to 8 assessment waves (*M* = 2.72 assessment waves), 14 years after posttest (range = 4 to 49 years after posttest; Figure S3).[3] The average posttest impact was .08 *SD* (*p* = .06) for social-emotional outcomes (74% of sample) and .24 *SD* (*p* = .006) for cognitive outcomes (26%). Impacts on adult outcomes were generally between .10 *SD* and .20 *SD* (range = .04 *SD* to .28 *SD*), split across the domains of substance use (28%), education (23%), psychological wellbeing (14%), employment (10%), cognitive functioning (6%), crime (6%), health (4%), social service receipt (1%), and other (10%).

Patterns of Longitudinal Intervention Impacts

Overall, interventions generated initial effects that faded considerably but nonetheless produced long-run adult impacts (Figure 3; Table S5). The observed pattern did not follow U-shaped depictions of the fadeout-emergence paradox (e.g., "Fadeout-Emergence" in Figure 1). Instead, long-run adult effects were similar in magnitude to faded-out medium-term effects.

Across all outcomes, we observed a posttest effect of .15 *SD* (*p* = .006) that faded to .10 *SD* (*p* = .04) at 6- to 12-month follow-up and .13 *SD* (*p* = .02) at 1- to 2-year follow-up, followed by an adult effect of .11 *SD* ($p < .001$). As Figure 3 reflects, we observed heterogeneity across studies at posttest ($\tau = .24$) and follow-up ($\tau = .11 - .15$; Figure S4). Collapsed by intervention, we observed a posttest effect of .14 *SD* (*p* = .003) that faded to .07 *SD* (*p* = .07) at 6- to 12-month follow-up and .06 *SD* (*p* = .08) at 1- to 2-year follow-up, followed by an adult impact of .04 *SD* (*p* = .04). As a result of aggregation, there was less between-study variation in effects.

Persistence in the Medium Term

---

[3] The minimum follow-up time elapsed, 4 years from posttest, is to be expected given that, to be included in our sample, adult follow-up could be measured as early as at the end of high school and interventions had to end prior to the beginning of high school.

To examine the rate of posttest persistence in the years after programs ended, we modeled the proportion of the initial effect that persisted at medium-term follow-up. Only 13 interventions reported impacts on the same outcome at follow-up. We observed a larger posttest effect of .20 *SD* (p = .09) among outcomes measured consistently, which faded to .15 *SD* (p = .05) at 6- to 12-month follow-up and .08 *SD* (p = .31) at 1- to 2-year follow-up. The average adult impact for studies with at least one consistent medium-term follow-up measure was similar in magnitude to that in the full sample: .10 *SD* (p = .15). Our regression models suggested that conditional on the posttest impact magnitude, effects persisted at a rate of about 50% (p = .05) 6 to 12 months after intervention end, and 25% at 1- to 2-year follow-up (p = .10; Table S6).

Forecasting Adult Follow-up Impacts using Posttest Impacts

We next tested whether initial intervention impacts on child outcomes forecasted long-run impacts on adult outcomes using impacts aggregated to the intervention level (Figure 4; Table S7, column 1). Our model estimated a statistically non-significant and imprecise association between posttest and adult impacts of β_1 = .13 (p = .30). If the observed .13 effect reflected a real causal impact, it would indicate that for every 1 *SD* increase in an intervention's average posttest effect size, we could expect a .13 *SD* increase in an intervention's average adult effect size. The model also produced a statistically non-significant intercept of .03 *SD* (p = .09). Taking the magnitude at face value suggests that a considerable portion of adult impacts (.03/.04 or 75%) were explained by factors not captured by child posttest impacts. However, the estimate was statistically non-significant and could reflect noise. Of note, in analyses that dropped each study (Table S8), we observed that the slope decreased from .13 to .04 (p = .68) when we dropped Perry Preschool and increased to .21 (p = .09) when we dropped Higher Achievement.

Publication Bias and Selective Reporting

We then examined issues of selective reporting and publication bias. One concern is that studies with adult effects observed larger-than-typical posttest or medium-term effects (Watts et al., 2019). However, compared with Hart et al. (2024), posttest impacts in our sample were actually smaller (Figure S5) and medium-term persistence rates were only slightly stronger.

Funnel plots showed a slight positive skew wherein larger positive effects were reported more than expected (Figure S6). A PET test indicated that studies with less precise estimates observed larger effects (Chen & Pustejovsky, 2025; Stanley & Doucouliagos, 2014) at posttest (β_1 = .39, p = .48), 6- to 12-month follow-up (β_1 = .59, p = .13), 1- to 2-year follow-up (β_1 = .52, p = .18), and adult follow-up (β_1 = .61, p = .02). PET-adjusted estimates (i.e., estimates when standard errors = 0) were 40 to 90% smaller than our non-adjusted estimates (Table S9). However, the PET model can underestimate non-zero effects and provides a crude estimate considering the substantive reasons why smaller studies may produce larger effects.

P-curve analyses did not suggest any major upticks in the number of p-values near the .05 conventional "statistical significance" cut-off (Figure S7; Simonsohn et al., 2015). The analysis did indicate that power to detect statistical significance was particularly compromised at the medium-term follow-up waves. Funnel plot and p-curve models using alternative standard errors estimated lower power for detecting adult effects (Table S10; Figures S8, S9). Selection models (Pustejovsky et al., 2025; Vevea & Woods, 2005) produced inconsistent results and together suggested that there did not appear to be egregious evidence of bias (Table S9; i.e., underreporting of small, statistically non-significant results). Of note, the PET model, one of the four selection models, produced an adult impact that was only 9% of the estimated effect.

Additional Analyses

The supplement details exploratory analyses that we will briefly review here. First, given similarities between the observed effects in our sample (Figure 3) and the "Network-Level Effects" theory in Figure 1, we examined simulated estimates based on the LINT theory (Bailey et al., 2024; Figure S10). The simulated estimates appeared similar to the observed effects. Critically, however, this study was not designed to test LINT nor to adjudicate across competing theories, so we interpret this analysis as hypothesis-generating rather than confirmatory.

Second, we fit additional models relying on small subsamples and, thus, caution against strong interpretation of the results. The association between posttest and adult effects was larger for cognitive skills than for social-emotional skills, though the difference was noisy and not statistically significant (Table S11; Figure S11). The correspondence between posttest and adult effects was smaller for educational adult outcomes than for other outcomes, and the slope for substance use outcomes was larger than for other outcomes (Table S11). Fully interacted models produced no statistically significant differences (Table S12). The association between posttest impacts and impacts on the most consistently reported adult outcome, high school attainment, was .36 (p = .10; Table S13, Figure S12). For the studies that measured them, medium-term follow-up effects were highly predictive of adult effects (Table S7, Figure S13). Adult assessment timing did not moderate the link between posttest and adult effects (Table S14).

We also performed many robustness checks to examine the sensitivity of our findings. We found that the results largely aligned with our primary findings across the following checks: alternative weighting and standard errors (Table S10); CHE model (Table S10; Pustejovsky & Tipton, 2021); controls for assessment/intervention timing (Table S15); dropping: high-attrition adult outcomes, studies that collected posttest measures before the end of the treatment, "lifetime" or "ever" measures (Table S16); alternative aggregation of adult effects (Table S16).

When we dropped effects that were more computationally intensive (e.g., categorical outcomes), the correspondence between posttest and adult effects increased (β_1 = .61, p = .06). However, upon dropping Perry Preschool, the slope decreased (β_1 = .37, p = .14). The slope was also larger in a model that linked average posttest impacts to individual adult impacts (β_1 = .32, p = .02). Finally, we applied Gilbert and Himmelsbach's (2026) error adjustments and found that accounting for error in posttest effects increased our slope term from β_1 = .13 to β_1= .14 (Table S17). As the assumed correlation between posttest and follow-up error increased, our slope estimate decreased. Slope terms ranged from β_1 = .10 to β_1 = .04 under r = .10 to .30.

Discussion

The mystery of sleeper effects—whether they are real and what they imply—has long captured the interest of social scientists (Table S1). In recent years, several rigorous educational intervention evaluations have found patterns that appear to fit the bill: long-run impacts on adult outcomes despite absent, or fading, intermediary impacts on child skills (e.g., Chetty et al., 2011; Deming, 2009; Gray-Lobe et al., 2023). These patterns, coupled with increasing evidence for the ubiquity of medium-term fadeout (Hart et al., 2024), raise important questions about how interventions affect life course outcomes, if not through persistent impacts on targeted skills.

The current paper first reviewed the literature on the life course effects of childhood interventions with a focus on theory, areas of complication, and methodological challenges. We then set out to conduct a broad examination of the longitudinal impacts of RCT-evaluated educational interventions. We first observed how widespread the fadeout-emergence pattern was. Necessarily, we were limited to RCTs that followed participants into adulthood, which was a relatively small and selected sample of 29 programs (Watts et al., 2019). Yet, we reasoned that by examining broad patterns of average impacts, our meta-analysis could improve the field's understanding of the general state of a literature that has been challenging to consolidate.

We did not find widespread evidence of the "fadeout-emergence paradox." Instead, we found that initial effects on child skills faded in the two years after programs ended, and programs produced statistically significant impacts on adult outcomes, the magnitude of which (.04 to .11 *SD* depending on the model) was similar to that of faded-out impacts on child skills. Adult impacts were similar in magnitude across a wide range of outcomes. Longitudinal program effects appeared to align with the hypotheses of LINT (Bailey et al., 2024), although the pattern could align with other theories, and this study did not aim to test competing theories.

Several features of our approach likely contributed to why we did not observe the pattern of fadeout-emergence as commonly construed, wherein large initial effects fade but are followed by large adult impacts. By representing all effect sizes in standard deviation units, we avoided misleading comparisons of impacts in raw units over time. We also computed intervention impacts on all posttest and adult outcomes, including those that were not explicitly featured by authors. By analyzing average impacts across outcomes, we avoided overweighting individual outcomes. Finally, we analyzed main effects rather than subgroup estimates, which may be more likely to suffer from reporting bias (von Hippel & Schuetze, 2025).

After establishing that programs can generate long-run effects despite medium-term fadeout, we probed the basic assumption that interventions with greater impacts on child skills also generate larger effects on adult outcomes. We examined the level of correspondence between impacts on child skills and long-run impacts on adult outcomes. In line with our descriptive observations, we found strong correspondence between medium-term impacts and long-run adult effects among the subsample of studies that reported medium-term follow-up.

However, we did not find much predictive value for posttest impacts. We observed a positive (non-statistically significant) association between posttest impacts on child skills and follow-up impacts in adulthood of β_1 = .13, but we also observed a .03 *SD* intercept (non-statistically significant), suggesting that some portion of the average .04 *SD* adult impact could be due to factors unexplained by post-test impacts. The intercept was imprecise, however, so we cannot estimate the exact portion. The slope was sensitive to the inclusion of two studies, Perry Preschool and Higher Achievement, and to some selection model adjustments (Table S8, S9).

Exploratory analyses revealed that the association between child skills and adult effects may vary depending on how they are operationalized. Posttest impacts on cognitive skills were more predictive of long-run outcomes than social-emotional skills were (β_1 = .16 versus β_1 = -.06), and we observed greater correspondence between average posttest impacts and impacts on the adult outcome measured most consistently: high school graduation (β_1 = .36).

Along with the finding of medium-term fadeout, the small association between posttest and adult effects suggests that it is highly unlikely that persistence on targeted child skills will necessarily produce adult outcomes. Skills may beget skills, but in complicated ways consistent with fadeout. At present, average posttest impacts on child skills do not appear to be a reliable and robust forecaster of long-run impacts on adult outcomes.

We do not view these results as cause for despair. Imagine that in a decade from now, with a larger sample of studies, we observe that the ~.10 association between child skills and adult outcomes is consistent and replicable, leading us to conclude that .10 is the causal long-run impact of changes to child skills targeted by interventions. Although far from a 1:1 association between targeted child skills and important adult outcomes, this would suggest that posttest impacts on child skills contain some meaningful information for forecasting long-run effects.

Certainly, our reliance on average impacts across studies may not satisfy readers who hold specific theories about why certain interventions might have stronger effects than others in the long term. We agree that complex combinations between skills and contexts may dictate impact persistence, and our analyses cannot capture this complexity. Indeed, our exploratory moderation analyses suggested that the overall association between posttest and adult effects could mask heterogeneity in addition to being limited by issues of statistical power and measurement error. However, related meta-analytic work has found that short-term persistence rates were largely unrelated to intervention characteristics (Watts et al., 2025). We encourage researchers who hold specific theories to test these with *a priori* predictions and experiments.

We were surprised to observe a stronger link between medium-term follow-up impacts and adult follow-up effects—ranging from 0 to .94 depending on our approach (average $\beta_1 =$ .42)—although our estimates relied on a small subsample and were imprecise. The associations indicated some promise for using medium-term follow-up effects to proxy program's network-level impact, which may correspond strongly with adult effects (see also Gilraine & Pope, 2021). Research is needed to determine when, if at all, medium-term follow-ups predict adult impacts.

Limitations and Recommendations for Future Research

Our experience creating the meta-analytic sample for this study revealed issues in the intervention evaluation literature that limited our results and warrant discussion.

First, we urge researchers to use administrative data to examine the long-run adult effects of existing RCTs regardless of posttest impact magnitude or medium-term fadeout. Reflective of the state of the field, our study was limited by a small sample of studies with relatively small sample sizes. More studies are necessary to examine heterogeneity; indeed, we could not adequately test study-, participant-, and outcome-related variation in the patterns we observed.

Second, we encourage researchers to measure the same outcomes across studies and assessment waves. It is possible that our operationalization of initial and long-run intervention impacts, which relied on intervention-level averages of all impacts, did not accurately capture "true" effects for a variety of idiosyncratic reasons such as over- or under-aligned to intervention content (Alvarez-Vargas et al., 2023) and selective measurement based on earlier results (Bailey et al., 2020). Establishing a standardized set of outcomes that intervention researchers consistently measure across studies and waves, regardless of intervention targets, may be fruitful.

Third, we call upon researchers to report all effects consistently and straightforwardly. Our results may be biased by problematic practices, including a) reporting subgroup effects in lieu of main effects, b) using inconsistent methods to compute impacts, and c) selectively reporting results in the socially preferred direction. We attempted to overcome these issues through our coding and analysis approach, but it is unlikely that our method overcame all biases.

Fourth, we recommend that researchers match the specificity of their analyses and conclusions to that of their hypotheses. Theoretical underdetermination is an issue that interacts unfavorably with inconsistent measurement and reporting, distorting our understanding of longitudinal intervention impacts. Researchers' theories of change can become increasingly distant from reality when anchored to potentially spurious results (e.g., subgroup estimates, selectively reported estimates). If researchers do not have specific theories about which adult outcomes their intervention will affect, it may be preferable to pre-register primary models that pool across measures. At a minimum, we recommend that researchers formalize research questions and analytic plans internally prior to analysis, as we did in the current meta-analysis.

Fifth, for researchers interested in causal skill-building processes, we advocate using observational methods only as part of a much richer research approach that capitalizes on exogenous variation in child skills and, whenever possible, probes the alignment between models fit to observational and (quasi-)experimental data. As detailed by Bailey et al. (2024), developmental theory suffers from an over-reliance on correlational evidence and an under-reliance on causal intervention evaluations. The disconnect between the inferences drawn from correlational studies and the causal findings from experiments demonstrates that correlational estimates can be misleading (Bailey et al., 2018; Rohrer & Lucas, 2023).

Finally, we emphasize that our data reflect the limitations of the current literature. Indeed, our sample was small, resulting in imprecise estimates. Although limited, our dataset reflects the current state of the field. We hope this work will provide researchers with a clearer picture of the current empirical basis for our theories on the long-term effects of educational programs targeting children's capabilities and sets the stage for future work on this topic.

Conclusion

The current study endeavored to synthesize the longitudinal effects of RCT-evaluated educational interventions targeting children’s skills. We found that educational interventions generated initial effects on child skills, these effects faded in the medium term, and programs nonetheless generated long-run impacts on adult outcomes. Adult impacts were generally similar in magnitude to faded-out medium-term effects. We observed some correspondence between initial intervention impacts on child skills and long-run impacts on adult outcomes, though our estimates were imprecise. There remains great opportunity for our field to continue to study how interventions shape life course trajectories. Theories that predict strong persistence of impacts on targeted child skills seem unlikely. We need more long-run evaluations of existing RCTs to make progress. Continued investigation can hone our understanding of development and improve the likelihood that interventions live up to their promise.

References

Alvarez-Vargas, D., Wan, S., Fuchs, L. S., Klein, A., & Bailey, D. H. (2023). Design and analytic features for reducing biases in skill-building intervention impact forecasts. *Journal of Research on Educational Effectiveness*, *16*(2), 271–299. https://doi.org/10.1080/19345747.2022.2093298

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association, 103*(484), 1481–1495. https://doi.org/10.1198/016214508000000841

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. *PSIP*, *21*(2), 55–97. https://doi.org/10.1177/1529100620915848

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

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

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

Blair, C., & Raver, C. C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review of Psychology*, *66*(1), 711–731. https://doi.org/10.1146/annurev-psych-010814-015221

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

Breit, M., Scherrer, V., Tucker-Drob, E. M., & Preckel, F. (2024). The stability of cognitive abilities: A meta-analytic review of longitudinal studies. *Psychological Bulletin*, *150*(4), 399–439. https://doi.org/10.1037/bul0000425

Brick, T. R., & Bailey, D. H. (2020). Rock the MIC: The matrix of implied causation, a tool for experimental design and model checking. *Advances in Methods and Practices in Psychological Science*, *3*(3), 286–299. https://doi.org/10.1177/2515245920922775

Bronfenbrenner, U. (1992). Ecological systems theory. In R. Vasta (Ed.), Six theories of child development: Revised formulations and current issues (pp. 187– 249). Kingsley Publishers.

Chen, M., & Pustejovsky, J. E. (2025). Adapting methods for correcting selective reporting bias in meta-analysis of dependent effect sizes. *Psychological Methods*, Online first. https://doi.org/10.1037/met0000773

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

Clarke, A. D. B., & Clarke, A. M. (1981). “Sleeper effects” in development: Fact or artifact? *Developmental Review*, *1*(4), 344–360. https://doi.org/10.1016/0273-2297(81)90030-7

Cook, T. D., Gruder, C. L., Hennigan, K. M., & Flay, B. R. (1979). History of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. *Psychological Bulletin*, *86*(4), 662–679. https://doi.org/10.1037/0033-2909.86.4.662

Cronbach, L. J. (1969). Heredity, environment, and educational policy. *Harvard Educational Review*, *39*(2), 338–347. https://doi.org/10.17763/haer.39.2.nvr226676j010551

Cunha, F., & Heckman, J. (2007). *The Technology of Skill Formation*.

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

Duckworth, A. L., Milkman, K. L., & Laibson, D. (2018). Beyond willpower: Strategies for reducing failures of self-control. *Psychological Science in the Public Interest*, *19*(3), 102–129. https://doi.org/10.1177/1529100618821893

Elango, S., García, J. L., & Heckman, J. J. (2016). *Early childhood education.* In R. A. Moffitt (Ed.), Economics of Means-Tested Transfer Programs in the United States (Vol. 2, pp. 235-297). University of Chicago Press. https://doi.org/10.3386/w21766.

Eronen, M. I. (2020). Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, *59*, 100785. https://doi.org/10.1016/j.newideapsych.2020.100785

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Freud, S. (1922). *Introductory Lectures On Psycho Analysis*. http://archive.org/details/in.ernet.dli.2015.278046

Gershkoff-Stowe, L., & Thelen, E. (2004). U-shaped changes in behavior: A dynamic systems perspective. *Journal of Cognition and Development, 5*(1), 11–36. https://doi.org/10.1207/s15327647jcd0501_2

Gilbert, J. B., & Himmelsbach, Z. (2026). *Why fadeout is (probably) worse than we think: adjusting for correlated sampling error in meta-analyses of behavioral interventions.* EdWorking Papers. https://doi.org/10.26300/87r9-qm15

Gray-Lobe, G., Pathak, P. A., & Walters, C. R. (2023). The long-term effects of universal preschool in Boston. *The Quarterly Journal of Economics*, *138*(1), 363–411. https://doi.org/10.1093/qje/qjac036

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

Heckman, J. J., Pinto, R., & Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. The American Economic Review, *103*(6), 2052–2086.

Hitt, C., McShane, M. Q., & Wolf, P. J. (2018). *Do impacts on test scores even matter? Lessons from long-run outcomes in school choice research.* American Enterprise Institute.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. (1949). *Experiments on mass communication. Vol. 3*. Princeton University Press.

Jones, D. E., Greenberg, M., & Crowley, M. (2015). Early social-emotional functioning and public health: The relationship between kindergarten social competence and future wellness. *American Journal of Public Health, 105*(11), 2283–2290. https://doi.org/10.2105/AJPH.2015.302630

Kagan, J. (1971). *Change and Continuity in Infancy.* New York: John Wiley & Sons, Inc.

Kellam, S. G., Ensminger, M. E., & Simon, M. B. (1980). Mental health in first grade and teenage drug, alcohol, and cigarette use. *Drug and alcohol dependence*, *5*(4), 273-304

Klahr, D. (1982). Nonmonotone assessment of monotone development: An information-processing analysis. In S. Strauss (Ed.), *U-shaped behavioral growth* (pp. 63–86).

Mak, M. H., & Twitchell, H. (2020). Evidence for preferential attachment: Words that are more well connected in semantic networks are better at acquiring new links in paired-associate learning. *Psychonomic Bulletin & Review*, *27*(5), 1059-1069. https://doi.org/10.3758/s13423-020-01773-0

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

Moffitt, T. E., Arseneault, L…. Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *PNAS*, *108*(7), 2693–2698. https://doi.org/10.1073/pnas.1010076108.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review, 26*(5), 1596–1618. doi.org/10.3758/s13423-019-016452

OpenAI. (2025). *ChatGPT (GPT-5)* [Large language model]. OpenAI. https://chat.openai.com

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/bmj.n71

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

Perc, M. (2014). The Matthew effect in empirical data. *Journal of the Royal Society Interface, 11*(98). https://doi.org/10.1098/rsif.2014.0378

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, *42*(8), 424-432. https://doi.org/10.3102/0013189X13507104

Poduska, J., Kellam, S., Wang, W., Brown, C. H., Ialongo, N., & Toyinbo, P. (2008). Impact of the Good Behavior Game, a universal classroom–based behavior intervention, on young adult service use for problems with emotions, behavior, or drugs or alcohol. *Drug and Alcohol Dependence*, *95*, S29–S44. https://doi.org/10.1016/j.drugalcdep.2007.10.009

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

Pustejovsky, J., Citkowicz, M., & Joshi, M. (2025). *Estimation and inference for step-function selection models in meta-analysis with dependent effects.* Metaarxiv. https://osf.io/preprints/metaarxiv/qg5x6_v1

Pustejovsky, J. E. (2023). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections.* R package version 0.5.10. https://CRAN.R-project.org/package=clubSandwich

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics, 36*(4), 672–683. https://doi.org/10.1080/07350015.2016.1247004

Pustejovsky, J., & Tipton, E. (2021). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*. https://doi.org/10.1007/s11121-021-01246-3

Rohrer, J., & Lucas, R. (2023). *Causal effects of well-being on health: It's complicated*. OSF. https://doi.org/10.31234/osf.io/wgbe4

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime effects: The High/Scope Perry Preschool study through age 40* (Monographs of the High/Scope Educational Research Foundation, No. 14). High/Scope Press.

Schweinhart, L. J., & Weikart, D. P. (1981). Effects of the Perry Preschool Program on youths through age 15. *Journal of the Division for Early Childhood*, *4*(1), 29–39.

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

Siegler, R. S. (2004). U-shaped interest in u-shaped development-and what it means. *Journal of Cognition and Development*, *5*(1), 1-10.

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.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*(1), 60–78. https://doi.org/10.1002/jrsm.1095

Thelen, E. (2005). Dynamic systems theory and the complexity of change. *Psychoanalytic* *dialogues*, *15*(2), 255-283. https://doi.org/10.1080/10481881509348831

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 428–443. https://doi.org/10.1037/1082-989X.10.4.428

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

van der Maas, H. L. J. (2024). *Complex Systems Research in Psychology*. Santa Fe Institute. https://santafeinstitute.github.io/ComplexPsych/

von Hippel, P. T., & Schuetze, B. A. (2025). How not to fool ourselves about heterogeneity of treatment effects. *Advances in Methods and Practices in Psychological Science*, *8*(2), 25152459241304347. https://doi.org/10.1177/25152459241304347

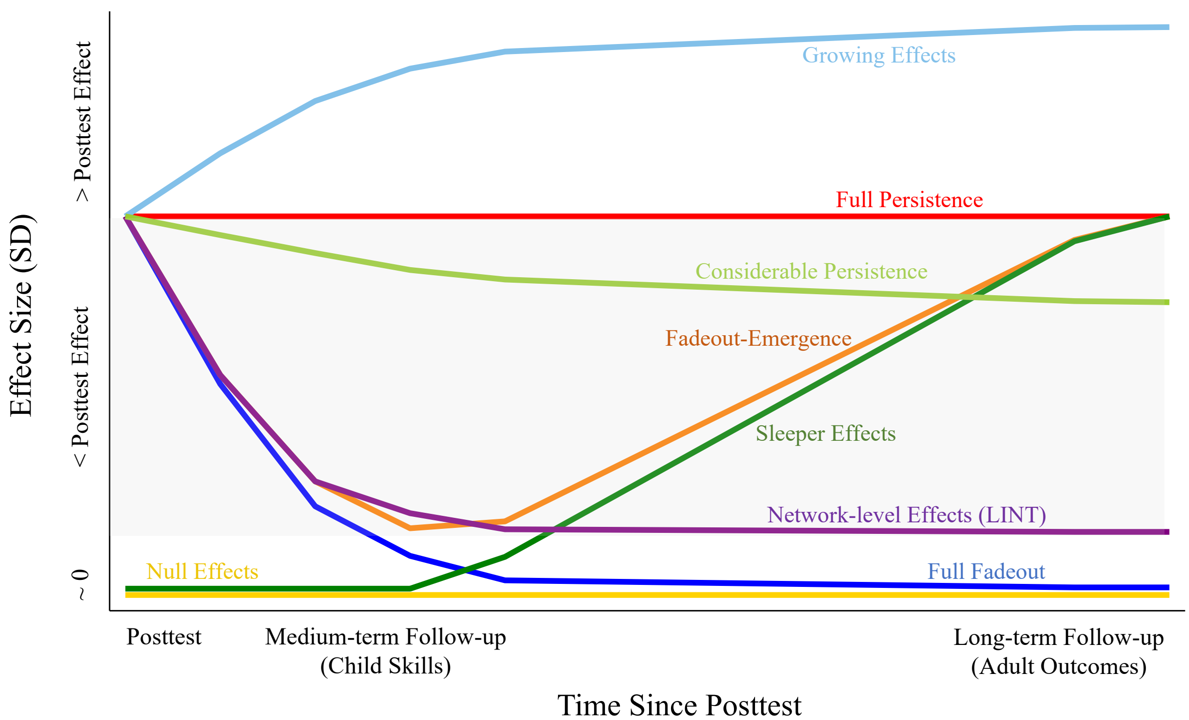
Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632-638. https://doi.org/10.1177/1745691612463078

Watts, T. W., Hart, E. R., Bailey, D. H. (2025). How general is educational intervention Fadeout? A meta-analysis of educational RCTs with follow-up. *Annenberg EdWorking Papers.* https://doi.org/10.26300/hxhd-cv70

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

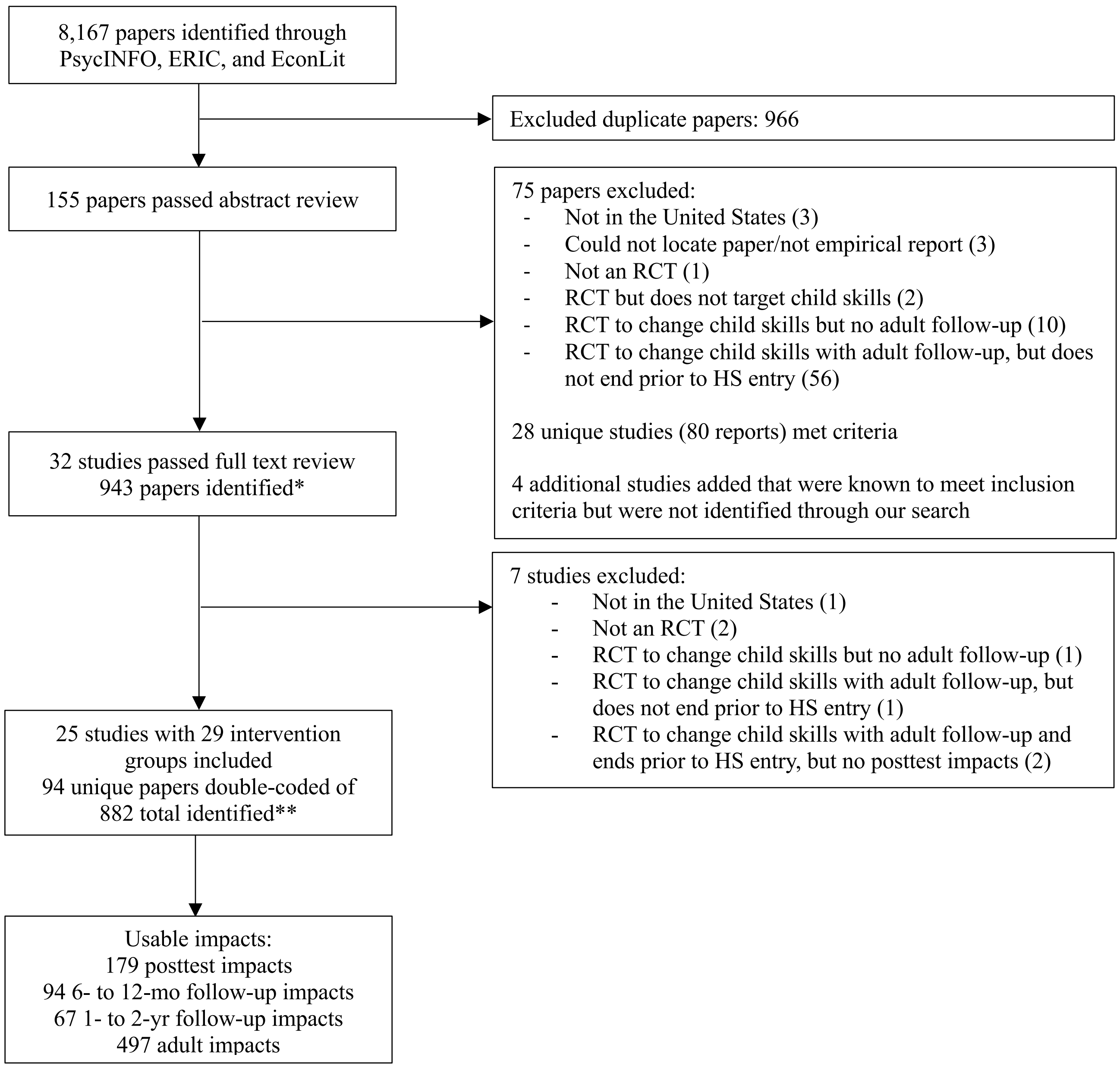
Werner, H. (1957). The concept of development from a comparative and organismic point of view.

Figure 1
Theoretical Trajectories of Intervention Impacts



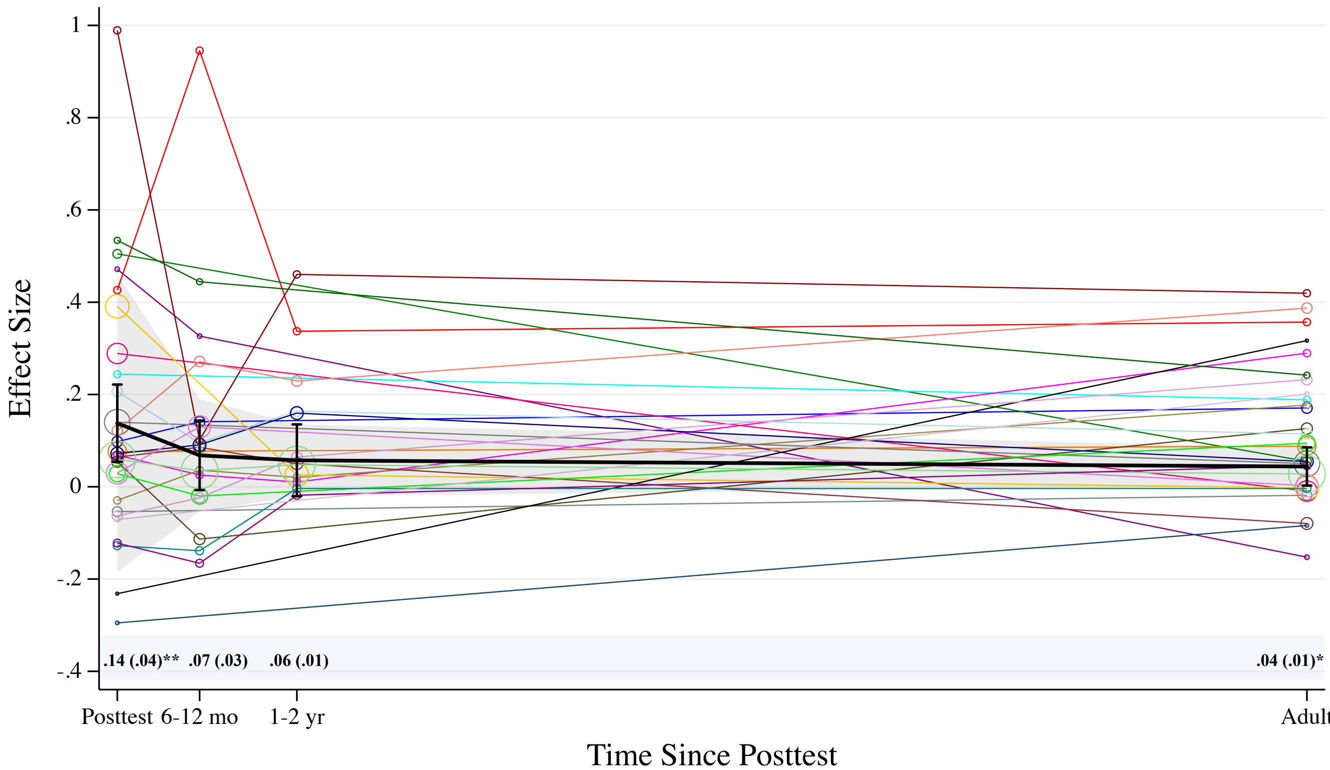
Notes: This figure depicts hypothetical trajectories of intervention impacts across time inspired by developmental theory. Note that the auto-regressive associations among child-outcome impacts across time and the cross-lagged associations linking child-outcome impacts to adult-outcome impacts are entirely hypothetical; developmental theories very rarely specify the predicted magnitude of these links.

Figure 2
Flow of Papers and Studies into the Meta-Analytic Sample



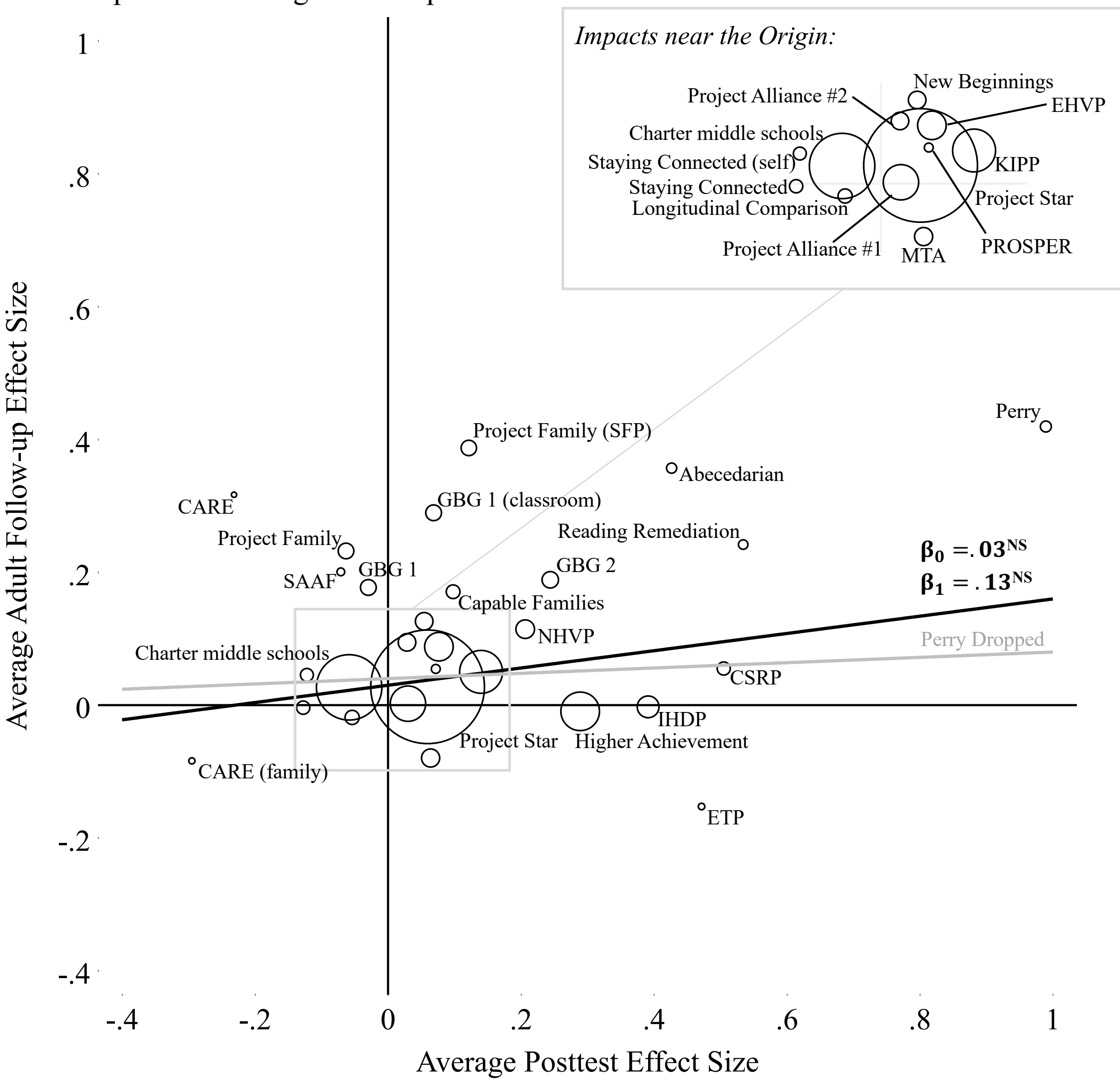
Notes: *For one study, we discontinued the search for additional papers after determining that the study was conducted outside the United States. **We identified 882 papers by summing the number of papers identified for each study. Because some papers reported results for multiple studies, they could appear more than once in this total. We ultimately coded 94 *unique* papers, corresponding to 101 potentially non-unique papers (i.e., 101 is the total when papers reporting results for multiple studies were counted once per study, as opposed to once across studies).

Figure 3
Trajectories of Average Effects for Each Intervention



Notes: Each line reflects one intervention. The coordinates reflect the intervention-level average impact for outcomes reported at the respective assessment wave. Coordinates are weighted by the posttest inverse sampling variances. The black line depicts the meta-analytic averages at each assessment wave, estimated using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). The black error bars reflect 95% cluster-robust confidence intervals. The gray shaded region reflects the 95% prediction intervals. For context, the red line with a spike at 6- to 12-month follow-up represents the Abecedarian study for which impacts on only one measure contributed (a measure of "conservation" using a Piaget-inspired task). See Table S5, Panel B for estimates.

Figure 4
Posttest Impacts Predicting Adult Impacts



Notes: This figure plots the average posttest effect size and the average adult follow-up effect for each intervention (n = 29). Coordinates are weighted by the follow-up inverse sampling variances. The black line depicts the estimated intercept (B_0) and slope (B_1) generated with weights for the follow-up inverse sampling variances, a random effect for study, and cluster-robust standard errors (with clustering at the study level). The gray line depicts the intercept and slope estimated using the same methods but with the Perry Preschool Intervention dropped. To reduce clutter, for studies with two intervention groups, the coordinate label without parentheses reflects the group that is not otherwise listed on the figure. See Table S3 for full intervention names associated with the label abbreviations.
* $p < .05$, ** $p < .01$, *** $p < .001$

Supplemental Information for:
Does Fadeout Preclude Long-Run Adult Effects? A Systematic Review and Meta-analysis of Educational Intervention Randomized Controlled Trials

Additional Methodological Details

Protocol

The protocol that guided the inclusion-exclusion process and coding will be posted publicly prior to publication.

Inclusion/Exclusion

Search Terms and Process

We determined the following search terms through an iterative process in which we tested various criteria for whether they led to the inclusion of studies that we knew should be included. We used the following search terms (AB = abstract; TI = title):

PsycINFO via EBSCO on 2/14/24

AB (adult* OR earn* OR economic OR achievement OR "high school graduation" OR college OR enroll* OR educational) AND AB (student* OR child* OR adolesc* OR infant* OR youth* OR grader* OR school*) AND ((AB(longitudinal OR long-term OR long-run OR follow-up) OR TI(longitudinal OR long-term OR long-run OR follow-up)) OR (TI(adult* OR earn* OR economic OR achievement OR "high school graduation" OR college OR enroll* OR educational) AND TI(randomly OR “randomized controlled trial” OR “randomized control trial” OR experiment OR experimental OR randomized OR lotter* OR effect*))) AND (AB(randomly OR “randomized controlled trial” OR “randomized control trial” OR experiment OR experimental OR randomized OR lotter*) OR TI(randomly OR “randomized controlled trial” OR “randomized control trial” OR experiment OR experimental OR randomized OR lotter*)) NOT AB(meta-analysis OR "systematic review" OR “nutrition” OR “natural experiment” OR “essays on” OR “lead exposure” OR chemotherapy OR cancer OR patients OR blood OR medical OR cardiac) NOT TI(meta-analysis OR "systematic review" OR “nutrition” OR “natural experiment” OR “essays on” OR “lead exposure” OR chemotherapy OR cancer OR patients OR blood OR medical OR cardiac)

Additional filters: Human population, English language, exclude dissertations
Results: 4,294

ERIC via EBSCO on 2/14/24

AB (adult* OR earn* OR economic OR achievement OR "high school graduation" OR college OR enroll* OR educational) AND AB (student* OR child* OR adolesc* OR infant* OR youth* OR grader* OR school*) AND ((AB(longitudinal OR long-term OR long-run OR follow-up) OR TI(longitudinal OR long-term OR long-run OR follow-up)) OR (TI(adult* OR earn* OR economic OR achievement OR "high school graduation" OR college OR enroll* OR educational) AND TI(randomly OR “randomized controlled trial” OR “randomized control trial” OR experiment OR experimental OR randomized OR lotter* OR effect*))) AND (AB(randomly OR “randomized controlled trial” OR “randomized control trial” OR experiment OR experimental OR randomized OR lotter*) OR TI(randomly OR “randomized controlled trial” OR “randomized control trial” OR experiment OR experimental OR randomized OR lotter*)) NOT AB(meta-analysis OR "systematic review" OR “nutrition” OR “natural experiment” OR “essays on” OR “lead exposure” OR chemotherapy OR cancer OR patients OR blood OR medical OR cardiac) NOT TI(meta-analysis OR "systematic review" OR “nutrition” OR “natural experiment” OR “essays on” OR “lead exposure” OR chemotherapy OR cancer OR patients OR blood OR medical OR cardiac)

Additional filters: English language
Results: 3,701

EconLit via EBSCO on 2/14/24

AB (adult* OR earn* OR economic OR achievement OR "high school graduation" OR college OR enroll* OR educational) AND AB (student* OR child* OR adolesc* OR infant* OR youth* OR grader* OR school*) AND ((AB(longitudinal OR long-term OR long-run OR follow-up) OR TI(longitudinal OR long-term OR long-run OR follow-up)) OR (TI(adult* OR earn* OR economic OR achievement OR "high school graduation" OR college OR enroll* OR educational) AND TI(randomly OR “randomized controlled trial” OR “randomized control trial” OR experiment OR experimental OR randomized OR lotter* OR effect*))) AND (AB(randomly OR “randomized controlled trial” OR “randomized control trial” OR experiment OR experimental OR randomized OR lotter*) OR TI(randomly OR “randomized controlled trial” OR “randomized control trial” OR experiment OR experimental OR randomized OR lotter*)) NOT AB(meta-analysis OR "systematic review" OR “nutrition” OR “natural experiment” OR “essays on” OR “lead exposure” OR chemotherapy OR cancer OR patients OR blood OR medical OR cardiac) NOT TI(meta-analysis OR "systematic review" OR “nutrition” OR “natural experiment” OR “essays on” OR “lead exposure” OR chemotherapy OR cancer OR patients OR blood OR medical OR cardiac)

Additional filters: Northern America (geographic region)
Results: 172

Restriction to Studies that Took Place in the United States

We decided to limit our search to interventions that took place inside the United States for several reasons. First, to our knowledge, the fadeout-emergence paradox has primarily been discussed in the context of U.S.-based interventions. Second, we anticipated that there would only be a small number of studies, if any, that met our criteria and were conducted outside of the United States (based on the findings of a recent meta-analysis by Hart et al., 2024). Third, we anticipated that the dynamics of fadeout, persistence, and emergence may vary across settings, particularly those with limited educational resources.

Abstract Review

Independent screeners reached consensus on all discrepancies. Peer-reviewed manuscripts, working papers, preprints, and government reports were included. At this stage, but not subsequent stages of the inclusion process, dissertations were excluded to reduce the review burden and given the low likelihood (given grant reporting requirements) that the only paper reporting adult follow-up effects for an intervention would be a dissertation.

Full-Text Review

We defined the two final inclusion criteria—that the intervention must end prior to high school and that adult outcomes must be measured after grade 12/age 18—after an informal assessment of the 155 papers screened at the full-text-review stage. We reviewed the papers with the intention of identifying criteria that would ensure that follow-up outcomes passed muster as measures of "adult" outcomes assessed in the "long-run," while also maintaining a sufficient number of studies for meta-analysis. Importantly, our informal review of the studies did not involve examining intervention effects, methodology, or any other details that could affect our results. We were focused, instead, on identifying criteria that had strong construct validity and allowed the inclusion of a sizable number of studies.

There were three interventions that we knew would meet our inclusion criteria that were included in our original MERF meta-analytic sample but were not identified through our screening process: the Infant Health and Development Program (McCormick et al., 2006), the Nurse Home Visitation Program (Eckenrode et al., 2010), and Project Care (Campbell et al., 2008). One additional study, Early Home Visiting (Conti et al., 2024), was identified through the subsequent forwards/backwards search process. Thus, while we are confident that we identified the vast majority of studies that met our criteria, that we did not find all known studies through our database search process raises the possibility that we may not have found *all* studies that met our criteria.

Backwards/Forwards Search Process

Our backwards and forwards search process entailed both searching within papers for additional papers on the intervention of focus and searching outside of papers (i.e., via Google Scholar). The search procedure is documented in detail in the protocol.

Choosing Papers to Code

When identifying papers to code, our priority was to choose a collection of papers that reported *all* posttest effects (i.e., end of treatment), medium-term follow-up effects occurring 6 months to 2 years after posttest, and adult follow-up impacts (i.e., collected after age 18 or grade 12) for each study. Although impacts were reported in traditional effect size units in many cases, we also included papers containing descriptive statistics (e.g., means and standard deviations) and model-based parameters (e.g., correlation coefficient, f-statistic) that we could then use to calculate an effect size in standard deviation units. When there were multiple papers that provided estimates on the same outcome, we chose to code the paper that provided the "best" effect (i.e., estimated on the larger sample size and with adjustments for issues like baseline imbalance, attrition, and clustered randomization).

Effect Size and Standard Error Computations

Figure S1 details how effect sizes were computed based on the reported information. When possible, we converted the author-reported statistics (e.g., odds ratio, predicted change in probability, hazard ratio) to Cohen's-d-like effects in standard deviation units. When only descriptive data (i.e., means, proportions) were reported, we used this information to calculate an estimate of the treatment-control difference. For descriptive dichotomous data, this entailed computing log odds ratios which we then scaled by $\frac{\sqrt{3}}{\pi}$. To standardize continuous treatment-control differences, we used the control-group standard deviations when available. If the control-group standard deviation was not available, we used other model-based parameters (e.g., F-statistic, confidence intervals, p-values) to estimate a standard deviation of the mean difference.

Valence. We adjusted each effect size so that all effects were scaled such that larger effect size indicated "better" or more desirable outcomes. We did this by multiplying each effect by an indication of whether the valence was positive (1) or negative (-1). We coded in a way that accounted for the measure's valence and whether the study-authors rescaled effects and/or descriptive outcomes. In coding valence, we opted to take each study on its own terms and at face value rather than ensuring that each outcome had a consistent valence code across studies. For example, if achievement of an associate's degree seemed to represent a desirable outcome in one study and a negative outcome in a different study, we coded valence as implied by the study authors. For some outcomes, the valence of an outcome was unclear. When it was not possible for us to form a reasonable inference about whether the outcome was considered a negative or positive outcome, we excluded the outcome.

Subsample Results. Our models only include estimates of main treatment effects. In the case that main treatment effects were reported, we coded these. In the case that only subgroup estimates were reported, and estimates were provided for all subgroups that together comprised the sample, we estimated main effects by computing a sample-size weighted average of the effect sizes for each group. Likewise, when computing the accompanying standard error, we generated a weighted average of the subgroup estimates.

Pre-test adjustments. As depicted in Figure S1, there were several cases in which we used author-reported statistics (e.g., *f*-statistics, *p*-values, standard errors, confidence intervals) to back out an estimated standard error which was then converted to an estimated standard deviation and subsequently used to standardize treatment and control group differences. In some cases, these statistics were estimated with controls for pre-test measures for a similar outcome, leading to increased precision. Following Hart et al. (2024), to ensure that this precision did not lead to an overestimation of intervention impacts, we divided the estimated standard error generated using the author-reported statistic by .87 for cases in which the author-reported statistic was estimated with a control for a pre-test measure similar to that of the outcome. The .87 represents the square root of 1 minus R^2, assuming a correlation between pre-test and posttest/follow-up assessments of .50. In many cases, our estimated effect sizes were subsequently used to calculate standard errors. We re-adjusted the standard errors associated with these effect sizes by multiplying them by .87.

Cluster adjustments. We adjusted all standard errors from studies using clustering by an intervention-specific VIF using the intervention-specific sample size (average posttest sample size for each intervention) and the number of clustering units, assuming an ICC of .10. For cases in which we used an author-reported standard error and the authors reported having already adjusted for clustering, we did not apply the cluster adjustment. Note that we did not do any computations to account for the fact that some effect sizes were computed using author-reported parameters (e.g., standard errors, *p*-values, confidence intervals) that contained cluster adjustments. When computing cluster-adjusted standard errors for our alternative standard errors, computed through more complex methods detailed in Figure S2, we did not perform the ICC adjustment if we estimated the standard errors using reported parameters (e.g., confidence intervals, *p*-values) that came from models that reportedly adjusted for clustered randomization.

Additional Details on the Dataset

What Kinds of Programs are in the Sample?

A diverse collection of educational programs has achieved the feat of measuring adult impacts. Here, we broadly categorize them.

Early Care and Education (ECE)

The first category comprises ECE studies, which were the first educational RCTs to measure impacts into adulthood, beginning in the 1960s. Programs like Perry and Abecedarian, in addition to the Early Training Project and Project Care, examined the impact of early classroom-based care for preschool-aged children, often coupled with additional family and health services. The Chicago School Readiness Project examined the effects of a specific curriculum in Head Start centers, marking the only early-childhood curricular intervention that measured adult impacts. A second set of programs—the Infant Health and Development Program, and two Nurse Home Visiting evaluations—targeted younger children, and centered home visiting (with center-based care in some cases).

Elementary-Level

In the elementary years, there were four curricular intervention evaluations that measured adult effects. Two interventions were focused on the development of academic skills, one involving reading remediation and the other instructional approaches for special education. Two other studies evaluated the effects of the Good Behavior Game, described earlier. Additionally, the Project Star experiment focused on class size and the Multimodal Treatment Study of ADHD entailed an evaluation of behavioral and pharmacological treatments (per our inclusion criteria, we focused only on the behavioral condition).

Middle School Transitions

Three programs focused on middle school as a critical transition. Two involved lottery evaluations of charter middle schools, including Knowledge Is Power Program (KIPP) and a national synthesis of 33 lotteries. A third program, Higher Achievement, involved intensive afterschool and summer programming for precocious students.

Middle School Positive Youth Development

The final programmatic category included positive youth development programs to promote prosociality and prevent problematic behaviors (e.g., substance use) in middle school. The programs in this category largely focused on supporting students to develop coping strategies and life skills. Often these programs also targeted parents. The programs included Project Family, two evaluations of Project Alliance, Capable Families and Youth Strengthening Families Program, PROSPER, and Staying Connected, many of which relied on similar, if not the same, underlying curricula. Two programs focused on specific populations, with intervention content tailored accordingly: New Beginnings, for students with divorced parents, and Strong African American Families, for Black students.

Additional Details on Primary Models

Aligned Outcome Models

In these models, we computed meta-analytic averages at each assessment wave only taking into account the interventions that measured at least one outcome consistently (using the same measure) at posttest and 6- to 12-month follow-up. For each intervention, we computed intervention-level averages at each assessment wave. Each posttest and 6- to 12-month follow-up intervention-level average only incorporated estimates from the outcomes that were consistently assessed at both waves. Each intervention-level average at 1- to 2-years follow-up included estimates from the outcomes that continued to be consistently assessed at this follow-up wave. Each intervention-level average at adult follow-up incorporated all adult outcomes that were assessed. In the case that there were multiple intervention impacts collected for the same outcome using the same measure within either the 6- to 12-month or 1- to 2-year follow-up windows (e.g., 6- and 9-month follow-up assessments), we computed an average for use in our analyses.

Correspondence between Posttest Impacts and Adult Follow-Up Impacts

There was no obvious way to link specific posttest measures to adult measures as we did with our aligned group approach. Indeed, on average, studies in our sample reported impacts on 6 child outcomes at posttest and 16 adult outcomes at follow-up yielding a multiplicity of child-skill-adult-outcome combinations.

Thus, in our primary model we relied on intervention-level averages. We acknowledge that researchers have highly sophisticated theories about the specific child skills that may affect specific adult outcomes. We do not doubt that, in many cases, these theories may accurately reflect the complexity of developmental processes. However, these theories are often vaguely articulated in published materials and, when articulated, it can be very challenging to gauge whether theories were based on *post hoc* or *a priori* reasoning. To overcome these concerns and given how little we understand about the mechanisms that shape long-run intervention impacts, extracting beyond study specificities by examining intervention-level averages was essential to the objectives of this study.

With that said, a possible critique of this approach is that we are 'comparing the incomparable' by examining average posttest and adult follow-up impacts from such a diverse, and relatively small ($k < 30$) collection of studies that each have their own specific theories of change. We agree that our approach does not incorporate study-specific nuance. However, we believe that examining average effects from a diverse set of studies is preferable to highly specified models, because it allows us to test the most basic assumption driving intervention work: that initial impacts on child skills relate to long-run intervention impacts on adult outcomes.

Random Effects

Four studies contained two randomly assigned treatment groups that were compared to the same control group ("treatment-control contrasts" or "interventions"; indeed, there were 25 studies and 29 interventions in our sample.) To account for these shared control groups, we opted to include study-level, rather than intervention-level, random effects.

Additional Details on Exploratory Analyses

LINT Simulation

Figure S10 depicts a simulated trajectory of intervention impacts based on LINT (Bailey et al., 2024). The simulated estimates came from two models in which we assumed that hypothetical interventions generated impacts on 4 skills measured at posttest. Auto-regressive paths on these skills were assumed to be .50 and cross-lagged paths between skills were assumed to be .10. The lower bound of the adult estimates (.02 *SD*) indicates the hypothesized adult impact under the assumption that a 1 unit increase in child skills caused a .20 unit increase in adult outcomes. The upper bound of the adult estimates (.05 *SD*) indicates the hypothesized adult impact under the assumption that a 1 unit increase to child skills caused a .30 unit increase in adult outcomes and that the intervention affected 4 non-measured outcomes in addition to the 4 measured skills.

Differences by Outcome Type

We then tested the extent to which the observed associations varied by posttest and adult outcome types (see Table S11). Indeed, it may be the case that some child skills are more causally influential in shaping some adult outcomes. Given the limitations of sample size and inconsistent assessment of child and adult outcomes across interventions, we examined outcome-related differences across broad outcome categories. We first examined whether cognitive versus social-emotional impacts were differentially predictive of average adult impacts (see Figure S11).

Next, we examined the correspondence between average posttest impacts and adult outcomes aggregated by the adult outcome categories with the most data coverage. In a subsequent step, we examined whether there were interactive and differential associations among cognitive versus social-emotional posttest impacts in predicting adult impacts in the three most represented adult outcome categories: educational outcomes, psychological wellbeing outcomes, and substance-related outcomes (see Table S12).

High-Occurrence Adult Outcome: High School Attainment

We then examined the association between average posttest impacts and the adult outcome that was reported most consistently across interventions: high school attainment (see Table S13 and Figure S12). Our sample contained 12 interventions that reported impacts on high school attainment. In the case that an intervention reported impacts on a high school completion outcome at multiple assessment waves, we went with the earliest assessment to increase cross-study harmony. As such, all estimates used in this analysis were collected when participants were 18 to 22 years old.

Medium-Term Follow-Up Impacts as Forecasters

Additionally, we examined the extent to which follow-up impacts measured at 6- to 12-month follow-up and 1- to 2-year follow-up forecasted adult impacts. We also limited the analysis to the subsample of "aligned" outcomes that were assessed consistently at posttest and medium-term follow-up to be able to make a direct comparison of differences in the associations with adult follow-up effects using the same sample. The results from these models are presented in Table S7 and Figure S13. We only had medium-term impacts from 18 interventions at 6- to 12-months follow-up and from only 13 interventions at 1- to 2-years follow-up.

Moderation by Adult Assessment Timing

Finally, we explored whether posttest impacts on child skills were better forecasters of more proximal adult impacts (see Table S14). Indeed, there was a great deal of variation in when adult effects were assessed, from when participants just entered adulthood to middle age.

Additional Details on Robustness Checks

Leave-One-Out Models

We ran models where we dropped each study to examine how study-level idiosyncrasies (such as those described in Table S3) may have biased our results (see Table S8).

Alternative Weighting and Correlated-and-Hierarchical Model

We then tested whether our results were robust to the use of alternative weighting approaches and the use of an alternative modeling approach: the Correlated-and-Hierarchical Effects Model (Table S10; Pustejovsky & Tipton, 2021), which accounts for within-study dependencies. Because the exact within-study dependencies among outcomes were unknown, we assumed that effects were correlated at $r = .60$ for the CHE model. Since we implemented CHE only as the working model and then clustered our standard errors using RVE, this approach was robust to misspecification of this exact dependence structure and value of the correlation. We opted to treat CHE as a robustness-check model given that our primary analyses had relatively few dependencies. Recall, we examined intervention-level averages from 29 interventions nested within 25 studies. To maximize consistency across our primary models and supplemental models, we opted to rely on the same–relatively parsimonious–primary model, at the expense of some models containing additional-and-unaccounted-for dependencies (e.g., the adult assessment timing model had multiple adult estimates from the same study).

Age and Timing Controls

Participant age at intervention implementation and at adult follow-up, and the time that elapsed between posttest and adult follow-up assessments, varied considerably across interventions in our sample. To address this, we ran our primary analytic model predicting follow-up impacts using posttest impacts with controls for three time-related variables: (a) the number of months elapsed between posttest and adult assessment, (b) participant age at adult assessment, and (c) participant age at intervention implementation as indexed by whether the intervention targeted children 7 years or younger or children older than 7 years (Table S15).

Attrition

For some adult outcomes there was considerable attrition. We computed the proportion of participants that contributed data for each adult outcome by dividing the total adult sample size for each outcome by the maximum sample size observed at posttest for the same intervention. After dropping outcomes for which retention was lower than 80% at follow-up (20% of outcomes), we then recomputed the intervention-level average adult impact and re-ran our primary model predicting adult impacts using posttest impacts (Table S16).

Pre-End-of-Treatment Posttests

As detailed in Table S1, there are some studies for which the posttest assessment was not a "true" posttest. Indeed, there were some cases where a study otherwise met our criteria, but there was no end-of-treatment assessment and, instead, a pre-end-of-treatment, intermediary assessment. We included these studies in our sample, but ran a model (see Table S16) that dropped studies with pre-end-of-treatment posttests.

Alternative Aggregation of Adult Effects to Address "Duplicate" Issue

When coding the posttest and medium-term follow-up effects, we were careful to ensure that we only coded one instantiation of the same measure at a particular time point. For example, if the full score and subscale scores for a particular measure were all reported at posttest, we opted to code the full score, and not the subscale scores. Thus, for each assessment wave, we only had one estimate per measure.

When coding adult effects, however, accounting for and avoiding duplicates was less straightforward. In many cases, effects were reported for various facets of one larger adult outcome. For example, "any educational attainment" "four year college attainment" and "two year college attainment" might all be reported. It was challenging to know in these cases what was the "best" and most comprehensive outcome to code. Thus, we opted to code *all* adult outcomes across all assessment waves, regardless of this issue. As a step to avoid our intervention-level adult impact averages being predominated by any one domain that had an outsized number of measures, we ran a supplemental model in which we computed the intervention-level average adult effect by first estimating the intervention-level average by construct domain (e.g., educational attainment, substance use, health) and then averaging these construct-level estimates (see Table S16).

Lifetime Measures

Another complicated issue was how to handle adult effects that were not time-limited. Consider, for example, a substance-use prevention intervention that measured drug use at posttest and measured "lifetime" drug use at follow-up. This adult effect could very well pick up on intervention-driven posttest differences. We attempted to indicate when a measure was a "lifetime" or "ever" measure when coding, though it was often challenging to make these distinctions. Nonetheless, we ran a model (see Table S16) that dropped the outcomes that we coded as "lifetime" or "ever." We know that this is an imperfect test because of the challenge of coding this distinction, and that we did not go to the lengths of determining for what studies and outcomes the "lifetime" status was a problem (e.g., the aforementioned drug use measure would not be an issue in the case of a preschool intervention, but would be an issue in the case of a substance use intervention).

Estimated Effects

As Figure S1 suggests, we employed a variety of approaches to estimating effect sizes. Some approaches relied heavily on estimation and computation. We ran a model (see Table S16) in which we dropped effects that were particularly heavily "estimated." These included: 1) impacts on dichotomous outcomes (e.g., odds ratios), 2) impacts that relied on an *t*-, *f*-, or p-statistic, and 3) impacts that relied on predicted changes in probability.

Alternative Specification: Individual Outcome-level Adult Impacts

To further probe whether outcome-level heterogeneity in adult outcomes was reduced after accounting for posttest impact magnitude, we then examined an alternative specification in which we linked each intervention-level average posttest impact to each individual outcome-level adult impact (Table S16). Importantly, this model contained challenging-to-account-for dependencies because intervention-level average posttest effects were represented multiple times, for each outcome-level adult effect.

Correlated Measurement Error

Error in posttest effects could bias our slope estimate down and correlated measurement error in posttest and adult effects could bias our slope estimate up. Gilbert and Himmelsbach (2026) recently proposed a new method that adjusts for measurement error in the predictor and allows for the estimation of the meta-regression intercept and slope terms under various assumptions about the extent to which measurement error in posttest and follow-up effects is correlated. The method employs Stan and several priors that are detailed in Gilbert and Himmelsbach's paper. We applied the method using our own base meta-regression model (random effects meta-regression with random effect for study and inverse variance weights). Like the authors, we ran the model under the assumption of r = 0, .10, .20, .30, .40, .50, .60, .70, .80, and .90. Results from the Stan models are presented in Table S17.

At r = 0, we can observe that the slope increases from .13 in our primary model to .14 in the adjusted model, suggesting slight disattenuation. We then examined estimated slope and intercept terms under different assumptions about the correlation between error in posttest and adult follow-up effects. First, we attempted to determine a reasonable back-of-the-envelope estimate of the average correlation between posttest and adult effects in our sample. We turned to two meta-analyses to form priors. First, a meta-analysis by Robson et al. (2020) examined the association between childhood social-emotional skills (specifically self-regulation) and a range of adult outcomes including substance use, mental health, and educational attainment. They found correlations between childhood social-emotional skills and adult outcomes that ranged from .05 to .22 (unweighted average = .14). Second, a meta-analysis by Strenze (2007) examined the association between measures of intelligence and academic performance and adult educational attainment, income, and occupation. They observed associations ranging from .11 to .46 (unweighted average = .33).

In our sample, 78% of the posttest effect sizes were social-emotional outcomes. Weighting the average estimate from Robson et al.'s self-regulation meta-analysis by .78 and the average estimate from Strenze's cognitive meta-analysis by .22, we arrived at an average correlation of r = .18. Of note, 35% of the effects in our sample were estimated with controls for baseline covariates. Assuming the use of baseline controls eliminates correlated error (which is probably a strong assumption), we might expect a correlation closer to r = .12 (.18*(1-.35)). Taken together, perhaps a correlation around r = .10 is a reasonable lower-bound estimate for the average correlated error and perhaps something closer to r = .30 would be a reasonable upper bound. A correlation of .10 corresponds with a slope estimate of .10, 77% of the estimate from our preferred model. A correlation of r = .30 corresponds with a slope estimate of .04, 30% of the estimate from our preferred model. Thus, the model indicates that the slope term may be lower than that produced by our preferred model under some specifications, but reasonable assumptions about the correlation between posttest and adult effects lead us to expect that the slope estimate should still be positive and greater than zero (although, not necessarily statistically different than zero).

It is worth noting a few caveats when interpreting the results from the Gilbert and Himmelsbach (2026) model. First, the model was originally employed to examine bias in cases when posttest and follow-up effects were measured using the same assessment over time. In our case, posttest measures of child skills are generally quite different from follow-up measures of adult outcomes. We anticipate that correlated measurement error should be less of an issue when measures change across assessment waves. Second, the model does not allow for heterogeneity in the assumed correlated error across studies and this may be problematic because error may be more correlated in some studies than others. Indeed, the child posttest outcomes and adult follow-up outcomes assessed across studies varied considerably and, as a result, there are likely some measures with error terms that are more correlated across time than others. Finally, some estimates were computed with controls for baseline covariates (35% of effects in our sample), whereas others were not. Adjustments for baseline controls should also significantly reduce concerns about correlated measurement error.

Table S1
References Describing the Puzzle of the Fadeout-Emergence "Paradox"

Reference	Quote
Dizikes, 2023	"It's not the case that we have an increase in test scores and it corresponds with an increase in college-going," Pathak says. "That's very intriguing." "If I had to speculate what's behind these long-term effects for college, this is our leading hypothesis," Pathak says of the reduction in behavioral problems. "There's a lot more that needs to be done on this. It's an intriguing finding. Others have highlighted these sorts of so-called noncognitive sleeper effects of education, and I've been quite skeptical about it. But now our own findings suggest there may be something to that story."
Farley, 2022	"There is an ongoing debate in early childhood about the phenomenon that test score impacts fade out, but effects re-emerge for adult outcomes," Parag Pathak, a professor in the economics department at the Massachusetts Institute of Technology and one of the authors of the controlled randomized trials in Boston, told us in an email. "This is sometimes called 'sleeper effects.' As a general rule, it is not a good idea to make decisions about effectiveness of education interventions only based on short-run outcomes. "We usually think that short-run academic outcomes are related to longer-term outcomes, but a growing body of literature argues that there are non-cognitive effects of education that are not easily measured via standardized assessments," Pathak said. "This is an active scholarly debate. My own view is that there is considerable uncertainty on both sides and we need more well controlled studies before any definitive statements can be made."
Chetty et al., 2011	"Studies in the broader literature usually find patterns like this: An excellent teacher or class can have a large effect on test scores in this year or the next, but most of the benefits have faded away within two or three years. The natural conclusion was, of course, that these effects must be only temporary and are unlikely to make a difference in the long run. We were surprised, then, to find a strong relationship re-emerge between kindergarten classroom quality and adult wage earnings!"
O'Donnell, 2010	"What's really surprising about our study," Chetty says, "is that [the advantage] comes back in adulthood."
Barnum, 2017	[DEMING] "I found that even though the impacts of Head Start faded out, there were long-run impacts on the things that we really care about, like high school graduation, college attendance, self-reported health, and then labor force participation, so whether you were working. The interesting thing about it was that it was actually the African-American kids and lower-income kids — kids whose mothers had lower academic achievement — they experienced the biggest benefits from Head Start and they had the most fade-out. It actually seemed like fade-out was somehow predicting longer impacts."

	[INTERVIEWER] "In the opposite of the way that you would expect?" [DEMING] "My lesson from that is I don't think we really know why test scores fade out, but in this case, I don't really care, because test scores are only useful to the extent that they correspond to something in a kid's life that we care about directly." [DEMING] "The bottom-line question is whether Head Start helps kids in the long run, and the answer is yes. My study is not the only one that finds that."
Gopnik, 2021	"Does preschool help children become more successful adults? Twenty-five years ago, I thought that there wouldn't be a straightforward answer to this question—development is just too complicated and hard to study. I was completely wrong. A clever and compelling study by the economists Guthrie Gray-Lobe, Parag Pathak and Christopher Walters has just appeared as a National Bureau of Economic Research working paper, and the answer is clearly 'yes.' But the results also raise fascinating questions about how early experience influences later life. In defense of my earlier view, development is, in fact, complicated and hard to study. How can we tell that there aren't other differences between children who go to preschool and those who don't? There are so many different kinds of preschool programs, so many different measures of success and so many other things that happen to children as they grow up. Wouldn't early effects just wash out later on? "Some of these problems have been reflected in policy debates. Early studies of big public programs, particularly Head Start, found that children's grades and test scores improved at first, but the effects faded out in elementary school. On the other hand, some more controlled and longer-term studies found that children who attended preschool were healthier and wealthier many years later. But these studies looked at experimental, relatively small interventions, like the Perry Preschool Program in Ypsilanti, Mich. In the new study, the researchers looked at both the short- and the long-term effects of a big public preschool program, and they took advantage of a remarkable natural experiment. In 1997, Boston decided to offer free universal preschool, but since there were more applicants than places, they used a lottery to decide who would get in. The result was the equivalent of a randomized, controlled trial—the scientific gold standard for determining cause and effect. The researchers analyzed the school records of over 4,000 students and compared children who had randomly won the lottery to those who hadn't. Most of the children were Black or Hispanic, and most were poor enough to be eligible for free lunch. The results are striking, and they actually make sense of the conflicting earlier results. As in the Head Start studies, the Boston preschools didn't influence short-term measures of school success, particularly standardized test scores in elementary and middle school. But, as in the Perry Preschool studies, there were long-term effects on other kinds of measures.

	"You'd have thought, as I once did, that preschool would have immediate effects on elementary school performance, influence high-school and so on, with each step leading to the next. Instead, we seem to see 'sleeper' effects on things like motivation and perseverance."
Von Hippel et al., 2025	"Where do these 'sleeper effects' in adulthood come from? No one really knows. Scholars used to talk vaguely about 'non-cognitive skills' that aren't measured by reading and math tests. Yet the Early Head Start Research and Evaluation Program measured a variety of social and emotional outcomes—including attentiveness, distractibility, peer relationships, 'externalizing' behaviors (acting out), and 'internalizing' behaviors (such as withdrawal or anxiety)—and found just as much fadeout as it did for the academic measures. At age five, the children in the Early Head Start group were doing better than the control-group children on many social and emotional outcomes, but by 5th grade the control group had caught up. Head Start Impact Study data show similar patterns, where short-term social-emotional benefits do not persist beyond the first year or two."
Slavin, 2017	"Long-term effects of effective programs are often seen, but how can there be long-term effects if there are no short-term effects on the way? Sleeper effects are so rare that many early childhood researchers have serious doubts about the validity of the long-term Perry Preschool findings."
Lehrer, 2010	"How does preschool work its magic? Interestingly, the Perry Preschool didn't lead to a lasting boost in IQ scores. While kids exposed to preschool got an initial bump in general intelligence, this dissipated by second grade. Instead, preschool seemed to improve performance on a variety of 'non-cognitive' abilities, such as self-control, persistence and grit. "
Petrilli, 2018	"We all hope that our preferred education reforms, at least the most ambitious ones, will 'change the trajectories of children's lives.' That's particularly true for children growing up in poverty, who typically face depressing odds of success if they attend mediocre schools. They may drop out before finishing high school, or they might graduate but with minimal skills. Either way, they're unlikely to complete postsecondary education or training or master the 'success sequence,' and are thus likely to work low-wage jobs and not enjoy the fruits of upward mobility. Many will struggle to form an intact family, and their children will grow up poor and then repeat the cycle. It's a bleak picture... "For all children, we hope that fantastic schools will benefit them in other long-term ways that are important but harder to measure: encouraging them to become active, informed citizens; identifying strengths and interests that they might put to good use in a career; and helping them become people of good character. "So, yes, long-term outcomes are extremely important, especially for children for whom positive outcomes are far from assured... "As the AEI authors write, it would be deeply disappointing if major reforms of schools and schooling moved the needle on test scores but didn't have a lasting impact on kids' lives…

	"Of course, it would be unfair to apply such a high standard to small, incremental changes, like adopting a better textbook or extending the school day. Nobody would expect tiny tweaks to have profound impacts, such as transforming a future high school dropout into a college graduate. But if they help lots of students become marginally more literate or numerate, they are still worth doing. "For major, expensive, disruptive reforms, however, stronger long-term outcomes are not too much to ask. And of course the most effective reformers are already aware of this. Consider KIPP, which has been obsessed from day one not just about student achievement but also at getting its KIPPsters to and through college, consistently tracking its numbers and refining its model. Which has properly led the organization to look at much more than just short-term test scores as indicators of whether their students are on track. Much of today's interest in non-cognitive skills grew out of this work."
Brooks-Gunn, 2003	Title: *Do You Believe in Magic?: What We Can Expect from Early Childhood Intervention Programs* "The large effects seen at the end of early education are not due to magic; they are based on what is known about young children's development, and the conditions and circumstances that promote or impede it. The ingredients of high quality early education are not magic, either, and may be repeated across centers, settings, populations, and regions of the country. To expect effects to be sustained throughout childhood and adolescence, at their initial high levels, in the absence of continued high quality schooling, however, is to believe in magic"
Ansari, 2019	"Finally, the sleeper-effects phenomena states that educational benefits of preschool may emerge later in the life course even in the absence of initial programmatic benefits (Clarke & Clarke, 1981). Although there has not been much empirical evidence for sleeper effects with classic evaluations of early childhood programs (Campbell et al., 2012; Schweinhart, 2006), which have documented fairly consistent impacts for children throughout the life course that carry forward (i.e., no break in the benefits of preschool), there has been growing recognition of the sleeper effects model in the developmental literature."

Table S2
References Interpreting and Applying Skill-Building Theories

Reference	Quote
Besser et al., 2025	"Instead, in line with the principle of self-productivity (Cunha & Heckman, 2007), which follows the idea that skills acquired in one period persist into future periods, and that skills are self-reinforcing, we assume that the initial boost that the children experience as a result of the teachers' participation in the training will persist over the course of the year between posttest and follow-up testing."
Garber et al., 2026	"Economic theories of building human capital (Cunha & Heckman, 2007, 2008 suggest a similar concept that 'skill begets skill.' ECE programs lay the foundation of skills on which children can continue to build skills during formal schooling. This theory suggests that ECE programs would not only give children an advantage early in elementary school, but that the effects of ECE programs would grow over time. However, continued growth of positive outcomes is often not the case in studies of ECE programs today (Abvenoli, 2019; ACF, 2012; Burchinal et al., 2024; Camilli et al., 2010; Lipsey et al., 2018)."
Bai et al., 2020	"Our framework is based on a dynamic model of skill formation (Cunha & Heckman, 2007; Cunha & Heckman, 2008). A child's skill is acquired through a multistage process. Inputs at each stage produce outputs at the next stage. Some skills may be more productively learned at some stages than other stages, and some investments may be more productive at some stages than at other stages. Skills acquired in one stage can persevere and afford learning of new skills at future stages. Cunha and Heckman (2007, 2008) call this perspective "skill begets skill" because basic skills are required to learn more advanced skills, such that trajectories of differences between initial-skill learners and non-learners will diverge over time. Initial intervention effects will not fade out but rather will grow. "Because skills developed in one period can persist and afford greater learning at future periods (Cunha & Heckman, 2007; Cunha & Heckman, 2008), we hypothesize persistent positive program effects for both SS and MF in their middle-school education outcomes."
King et al., 2023	"When viewed through the conceptual lens of developmental cascades, early competence can generate further well-being, such that positive functioning spreads to other domains over time (Masten & Cicchetti, 2010; Masten & Tellegen, 2012). Generally, however, our intent-to treat analyses did not reveal cascading positive effects of the foster care intervention. Instead, with the exception of externalizing problems, for which we observed a sleeper effect consistent with previous analyses in subsets of these data (Humphreys et al., 2015, Wade et al., 2023), we found that the benefits of foster care were similar throughout development."
Padilla-Walker et al., 2020	"We expected that over several years, engagement in prosocial behavior would initiate a positive developmental cascade of personal characteristics (hope, persistence, gratitude, and self-esteem), which would subsequently protect against the extent to which adolescents experience depressive and anxiety symptoms. These postulates are consistent with past research (Masten et al., 2005; Obradović et al., 2009) and the Developmental Cascades framework (Masten & Cicchetti, 2010), which suggest that frequently it is involvement in or experience with one factor

	(e.g., prosocial behavior) that initiates a positive cascade (e.g., healthy character traits), which in turn affects later adjustment (e.g., decreased depressive and anxiety symptoms). Therefore, overall competence in two domains (i.e., behaviors, personal characteristics) should beget competence in another (i.e., a reduction in internalizing symptoms (Lerner et al., 2015; Masten & Cicchetti, 2010, p. 492)."
Wolchik et al., 2025	"A central scientific goal of developmental psychopathology is to unravel the developmental pathways that lead to positive and negative adaptation outcomes (e.g., Cicchetti & Sroufe, 2000). One approach to studying these pathways is to model cascading effects, or the progressive associations among various domains of functioning over time (Masten & Cicchetti, 2010; Rutter & Sroufe, 2000). In these models, change in a particular area of functioning is viewed as initiating a domino effect of consequences, impacting both the initial area of adaptation and other areas of adaptation in subsequent developmental stages (Sameroff, 2000). Studies with the New Beginnings Program data set found support for cascade effects models in which program-induced improvements in positive parenting led to decreases in offspring's internalizing and externalizing problems in late childhood/early adolescence, which then led to lower mental health and substance use problems and higher work, academic, and peer competencies in adolescence, which in turn led to more adaptive functioning in emerging adulthood (Wolchik et al., 2016; 2021)."
Morawska et al., 2019	"Early self-regulatory processes provide a continually evolving framework, with more advanced self-regulatory processes building on mechanisms that develop earlier and thus gradually creating more sophisticated behavior as children mature (Eisenberg et al. 2010; Masten and Cicchetti 2010; Williams and Berthelsen 2017). For example, better self-regulatory capacity provides a framework for children to engage more effectively with teachers and peers in the early learning context, which in turn contributes to improved academic outcomes, fostering a virtuous cycle for continued improvement (Eisenberg et al. 2010; Williams and Berthelsen 2017)."
Zhang et al., 2023	"Because of the importance of understanding an intervention's targets that are responsible for its positive effects (Glenn et al., 2015), intervention studies with long-term follow-ups of youth provide an opportunity to investigate the developmental cascading pathways (Masten & Cicchetti, 2010) through which the targeted behaviors that were modified by the intervention can lead to a chain of positive effects across domains or systems over the course of development."

Table S3
Descriptives of the Included Interventions

Study ~ Treatment	Start year	Post ES (SE)	Adult ES (SE)	Avg *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies)*
Early Training Project (ETP) (Anderson, 2008; Klaus & Gray, 1968)	1963	0.47 (0.33)	-0.15 (0.39)	56	30	42	x	x	x	Improve the "aptitudes and attitudes" of children from extremely disadvantaged families in preparation for elementary school entry through summer preschool programming and year-round home visiting.	We combined the 3-year and 2-year groups given that some papers reported effects for the groups together. In some cases, we reached different conclusions about program impacts than the authors because we combined across subgroups, whereas the authors often focused on moderation by gender.
Perry Preschool Project (Elango et al., 2015; García & Heckman, 2023; Weikart et al., 1970)	1964	0.99 (0.21)	0.42 (0.23)	100	20	42	x	x	x	Improve the cognitive development of children with low IQ scores from extremely disadvantaged families prior to school entry through center-based care and home visiting.	In some cases we reached different conclusions about program impacts than the authors because we combined across subgroups, whereas the authors often focused on moderation by gender.
Abecedarian Preschool Project (Campbell & Ramey, 1990; Elango et al., 2015; García & Heckman, 2023; Pages et al., 2022)	1974	0.43 (0.21)	0.36 (0.25)	89	60	0		x	x	Improve the cognitive development of children from "high-risk" families prior to school entry through intensive center-based care, medical services, and nutrition support.	We focused on randomization to the initial early childhood intervention (rather than re-randomization to elementary intervention) given that this was the treatment of focus in adult follow-up papers. In some cases we reached different conclusions about program impacts than the authors because we combined across subgroups, whereas the authors often focused on moderation by gender.
Nurse Home Visitation Program (NHVP) (Eckenrode et al., 2010; Olds	1979	0.21 (0.15)	0.11 (0.14)	193	28	prenat al	x	x	x	Reduce child developmental problems in families with young, unmarried, and/or low-SES mothers through nurse home visiting targeting maternal wellbeing,	We used the pregnancy and infancy group as our treatment group of focus. The pregnancy-only group was not included because there are no posttest measures of child skills. The authors reported a deviation from "pure" random assignment: exclusion of 46 "non-white" participants.

Study ~ Treatment	Start year	Post ES (SE)	Adult ES (SE)	Avg *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies*)
et al., 1986, 1994)										parenting (e.g., maltreatment), and economic self-sufficiency.	
Project Carolina Approach to Responsive Education (CARE) ~ Center-based plus Family Education Group (Campbell et al., 2008, 2013)	1979	-0.23 (0.50)	0.32 (0.48)	33	96	0	x	x	x	Improve the cognitive development of children from "high-risk" families through intensive center-based care, parent coaching that emphasizes problem-solving skills and engagement in developmentally supportive caregiving behaviors, as well as high-quality elementary schooling through third grade.	In some cases we reached different conclusions about early intervention effects than the authors because the authors often disregarded the elementary-portion of the intervention in their discussion of these specific effects, whereas we treat end-of-third-grade impacts as the end of the intervention. Additionally, in some cases, we reached different conclusions about program impacts than the authors because we combined across subgroups, whereas the authors often focused on moderation by gender.
Project Carolina Approach to Responsive Education (CARE) ~ Family Education Group (Campbell et al., 2008, 2013)	1979	-0.30 (0.43)	-0.08 (0.41)	44	96	0	x	x	x	Improve the cognitive development of children from "high-risk" families through parent coaching that emphasizes problem-solving skills and engagement in developmentally supportive caregiving behaviors, as well as high-quality elementary schooling through third grade.	Our conclusions were very similar to those of the authors for this treatment arm, given that the treatment generally produced non-statistically significant impacts regardless of moderation or estimation approach (which were used in this group much like in the center-based plus family education group, described above).
Longitudinal Comparison Project (Mediated Learning)	1983	-0.05 (0.16)	-0.02 (0.18)	164	20	58		x	x	Improve learning outcomes of children who qualify for special education through child-led and scaffolded learning opportunities within the classroom to develop and apply domain-general cognitive and social skills (in contrast with academically oriented instruction).	This intervention compared two interventions: Mediated Learning and Direct Instruction. We treated Mediated Learning as the "treatment" group and Direct Instruction as the "control" group given that Mediated Learning appeared to be the more novel program. The authors noted that the average time students spent in the intervention was

Study ~ Treatment	Start year	Post ES (SE)	Adult ES (SE)	Avg *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies*)
(Cole et al., 1993; Jenkins et al., 2006)											1.65 years (presumably school years), and that the modal time spent was 1 year. The authors also noted that "posttest" assessment occurred 8 months after pre-test (approximately 1 school year on average). We included the author-defined "posttests" even though these appear to have been collected prior to intervention end for some participants. At least for the posttest analyses, the authors reported a deviation from "pure" random assignment: exclusion of participants who had previously attended preschool (excluded *n* is not reported).
Infant Health and Development Program (IHDP) (Hill et al., 2003; IHDP, 1990; McCormick et al., 2006)	1985	0.39 (0.07)	0.00 (0.12)	1016	36	0	x	x	x	Prevent behavioral, cognitive and health problems among low-birthweight premature infants through center-based care, home visiting, and parenting support groups.	We reached different conclusions about program impacts than the authors because we combined across subgroups, whereas the authors often focused on moderation by birthweight.
Project Star (Chetty et al., 2010; Dynarski et al., 2013; Kreuger, 1999; Muenning et al., 2011; Schanzenbach, 2006; Wilde et al., 2011)	1985	0.06 (0.04)	0.03 (0.02)	4073	44	65		x	x	Improve student performance through increased teacher-student ratios in early elementary school.	In some cases we reached different conclusions about early intervention effects than the authors because we regarded the end-of-third grade impacts (when intervention formally ended) as the end of the intervention. Estimates were generated using regular classrooms as the control group or using the regular classrooms plus regular classrooms with aides as the control group. We included impacts generated from both control group specifications. We reached different conclusions about program impacts than the authors because we included the

Study ~ Treatment	Start year	Post ES (SE)	Adult ES (SE)	Avg *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies*)
											social-emotional outcomes measured at posttest in addition to test scores.
Good Behavior Game 2 (GBG 2) (Dolan et al., 1993; Kellan, 2008; Poduska et al., 2008)	1986	0.24 (0.22)	0.19 (0.15)	289	20	72		x		Prevent psychopathology and antisocial behaviors through a universal classroom-based behavioral management program designed to prevent disruptive behaviors via peer-group reinforcement and reward system.	Impacts are for the Good Behavior Game group. The Mastery Learning group was not included because no adult impacts were presented. Across assessment waves, we reached different conclusions about program impacts than the authors because we combined across subgroups, whereas the authors often reported effects by moderators including cohort and gender. Importantly, the posttest assessments that we identified appeared to have several internal validity issues. First, we could only locate impacts for Cohort 1 and not Cohort 2. (For the adult follow-up, impacts were reported for both cohorts, which we combined for analysis.) Second, impacts were only reported for those within Cohort 1 who stayed in the condition for at least a year; hence, posttest impacts approximate TOT versus ITT estimates. Third, we could not locate "true" end-of-treatment posttest data for this intervention. Thus, we used impacts from the assessment wave closest to the end of the intervention which were collected 8 months into the intervention, a year before true intervention end.
Early Home Visiting Program (EHVP) (Conti et al., 2024; Donelan-McCall et al.,	1991	0.08 (0.08)	0.09 (0.09)	671	27	prenatal	x	x	x	Improve pregnancy outcomes and child development for "at-risk" families through home visiting services that provided detailed health and wellbeing support across pregnancy, education on caring for the health and development of young children, and support with life planning and problem solving.	Impacts are for the Pregnancy and Postnatal Group. The pregnancy group was not included because there are no posttest measures of child skills. When adult outcomes were presented separately by participant sex, we combined these.

Study ~ Treatment	Start year	Post ES (SE)	Adult ES (SE)	Avg *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies)*
2021; Kitzman et al., 1997, 2019; Olds et al., 2014)											
Multimodal Treatment Study of ADHD (MTA) (Hechtman et al., 2005; Jensen et al., 2007; Langberg et al., 2010; MTA, 1999, 2004; Swanson et al., 2017; Wells et al., 2006)	1992	0.06 (0.13)	-0.08 (0.14)	245	14	101	x	x		Improve symptoms of children with ADHD through multipronged behavioral intervention including an intensive therapeutic summer camp, teacher consultation and parent training by therapist, and part-time para-educator in classrooms.	Impacts are for the Behavioral Treatment group. We excluded the medication only and medication plus behavioral groups because they did not fit our conceptualization of an educational intervention.
New Beginnings ~ Mother-only and dual-component groups (Mahrer et al., 2014; Rhodes et al., 2019; Sigal et al., 2012; Vélez et al., 2011; Wolchick et al., 2013, 2021)	1992	0.05 (0.14)	0.13 (0.14)	240	3	124	x	x		Prevent negative effects of divorce on children (e.g., substance use and social-emotional problems) through sessions targeting child coping strategies, positive parenting behaviors, and mothers' management of conflict.	Some papers reported impacts in the context of complex SEM models. As such, in some cases we had to estimate effects from descriptive information.

Study ~ Treatment	Start year	Post ES (SE)	Adult ES (SE)	Avg *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies)*
Good Behavior Game (GBG)~ Classroom (Bradshaw et al., 2009; Ialongo et al., 1999; Musci et al., 2018; Wang et al., 2012; Saunders, 2007)	1993	0.07 (0.21)	0.29 (0.16)	403	8	74		x	x	Prevent psychopathology and antisocial behaviors through a universal classroom-based enriched curriculum and behavioral management program designed to prevent disruptive behaviors through peer-group reinforcement and reward system.	In some cases we reached different conclusions about program impacts than the authors because we combined across subgroups, whereas the authors often reported effects by moderators including gender. We estimated some impacts from papers that were not focused specifically on identifying intervention impacts.
Good Behavior Game (GBG)~ Family (Bradshaw et al., 2009; Ialongo et al., 1999; Musci et al., 2018; Saunders et al., 2007; Wang et al., 2012)	1993	-0.03 (0.21)	0.18 (0.16)	403	8	75	x	x	x	Prevent psychopathology and antisocial behaviors through universal program to strengthen parent-teacher partnerships and parents' engagement in at-home enrichment to support academic and behavioral development.	In some cases we reached different conclusions about program impacts than the authors because we combined across subgroups, whereas the authors often reported effects by moderators including gender. We estimated some impacts from papers that were not focused specifically on identifying intervention impacts.
Project Family ~ Preparing for the Drug-Free Years Group (Mason et al., 2003, 2007, 2009; Park et al., 2000; Spoth et	1993	-0.06 (0.15)	0.23 (0.16)	363	1	136	x	x		Reduce the risk of early substance use initiation through a universal program to strengthen parent-child relationships, parents' understanding of risk factors that shape substance use, and parents' strategies for supporting youths' positive behaviors.	The intervention was estimated to be approximately one month but, by our estimation, posttest effects were collected 6 to 9 months after program end (6 and 9 reported in different papers) thus posttest assessments were delayed. We estimated some impacts from papers that were not focused specifically on identifying intervention impacts.

Study ~ Treatment	Start year	Post ES (SE)	Adult ES (SE)	Avg *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies*)
al., 2006, 2008a, 2009a)											
Project Family ~ Strengthening Families Group (Spoth et al., 1999, 2002, 2006, 2008a, 2008b, 2009a, 2009b, 2013; Trudeau et al., 2012)	1993	0.12 (0.15)	0.39 (0.16)	374	2	136	x	x		Prevent substance use and other problem behaviors through a universal program for parents and children to strengthen their relationships, children's life skills and substance resistance strategies, and parents' strategies for supporting youths' positive behaviors.	The intervention was estimated to be approximately two months but, by our estimation, posttest effects were collected 6 to 9 months after program end (6 and 9 reported in different papers) thus posttest assessments were delayed. We estimated some impacts from papers that were not focused specifically on identifying intervention impacts.
Project Alliance #1 (Gardner et al., 2008; Panza, 2019; Van Ryzin et al., 2013; Van Ryzin & Dishion, 2012; Zhang et al., 2024)	1997	0.03 (0.07)	0.00 (0.07)	829	20	147	x	x		Prevent adolescent problem behaviors through a universal ecological intervention for parents with multiple levels including light-touch family resource center, parent-focused intervention that appraises child-level risks and advises on appropriate responsive parenting behaviors, as well as referrals for additional intervention.	While the authors stated in many papers that the intervention ended prior to high school, they note at least once that Family Check-Up was offered in high school. We treated eighth grade as posttest nonetheless. We estimated all impacts from correlation matrices reported in papers that either used complex SEM models or were not specifically focused on identifying intervention impacts, hence differences in the statistical significance and magnitude of author-reported effects versus our effects.
Reading Remediation (Blachman et al., 2004, 2014)	1998	0.53 (0.25)	0.24 (0.27)	69	8	95			x	Improve reading skills of students with poor word-learning skills through daily, individualized, one-on-one tutoring in “phonologic and orthographic connections in words and text-based reading.”	No major distinctions.
Capable Families and Youth Strengthening	1999	0.10 (0.15)	0.17 (0.18)	137 2	16	148	x	x		Prevent early substance use initiation through universal program for parents and children to strengthen parent-child	In some cases we reached different conclusions about early intervention effects than the authors because we treated the end of eighth grade (post-booster sessions) as the

Study ~ Treatment	Start year	Post ES (SE)	Adult ES (SE)	Avg *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies*)
Families Program (Spoth et al., 2002, 2005, 2006, 2008, 2013; Trudeau et al., 2003, 2016)										relationships, children's life skills and substance resistance strategies, and parents' strategies for supporting youths' positive behaviors.	end of intervention, whereas the authors treated the end of seventh grade as the end of the intervention. While not reported consistently across papers, the authors did note that there were additional booster sessions for a subset of students in the treatment groups in eleventh grade. We treated eighth grade as posttest nonetheless. We combined the Parents & Youth plus Life Skills Training group and Life Skills Training group because some papers reported effects for the groups together. We estimated some impacts from papers that were not focused specifically on identifying intervention impacts.
PROSPER Family- and School-focused Intervention (Redmond et al., 2009; Spoth et al., 2011, 2013, 2017, 2022)	2003	0.07 (0.12)	0.05 (0.28)	101 70	28	136	x	x		Prevent substance use and problem behaviors through community-level universal intervention systems for parents and children to strengthen parent-child relationships, children's life skills and substance resistance strategies and parents' strategies for supporting youths' positive behaviors.	In some cases we reached different conclusions about early intervention effects than the authors because we treated the end of eighth grade (post-booster sessions) as the end of intervention, whereas the authors reported and discussed effects assessed at earlier waves.
Chicago School Readiness Project (CSRP) (Raver et al., 2009, 2011; Watts et al., 2023)	2005	0.50 (0.18)	0.06 (0.19)	463	8	52		x		Improve school readiness, particularly emotional and behavioral adjustment, among children from disadvantaged families through training program to improve Head Start teachers' capacities to manage behavioral dysregulation including one-on-one support for teachers from mental health consultants.	No major distinctions.

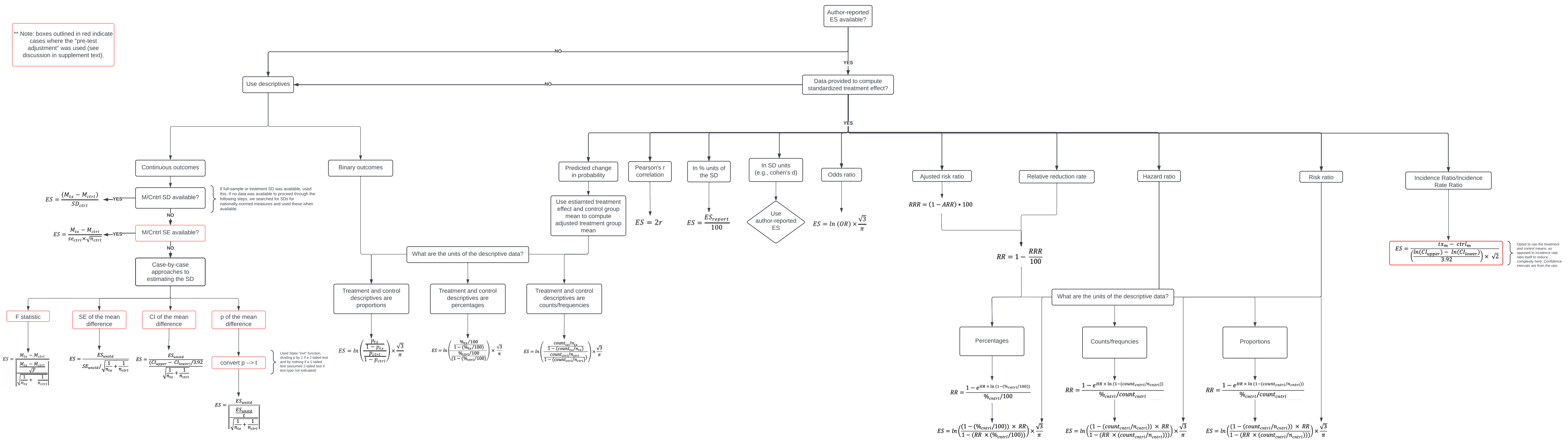
Study ~ Treatment	Start year	Post ES (SE)	Adult ES (SE)	Avg *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies)*
Charter Middle Schools (Clark et al., 2015; NCEE, 2019)	2006	-0.06 (0.04)	0.03 (0.04)	203 0	30	138		x	x	Improve middle school students' learning through charter schools with freedom and flexibility to innovate.	No major distinctions.
Higher Achievement (Herrera et al., 2013; Garcia et al, 2020)	2007	0.29 (0.08)	-0.01 (0.07)	719	42	118	x	x	x	Set motivated students from under-resourced schools on a college-going trajectory through afterschool and summer programming that supports academic achievement, social-emotional skill development, and application to high-quality high schools.	In some cases we reached different conclusions about early intervention effects than the authors because the authors focused on intermediary treatment effects estimated while the intervention was still ongoing, whereas we only focus on posttest effects. Of note, we did not include the high school attendance-related outcomes in our sample because they did not meet our criteria for cognitive or social-emotional outcomes.
Knowledge is Power Program (KIPP) (Coen et al., 2019; Demers et al., 2023; Tuttle et al., 2013)	2009	0.14 (0.06)	0.05 (0.06)	590	38	126	x	x	x	Improve academic achievement, and ultimately high school graduation and college attendance, for students from families with low incomes through a high-expectations charter school that seeks to increase children and families' engagement (i.e., time and effort) in the educational process.	"True" posttest data collected after intervention end was not provided. Thus, we considered the Terranova assessments, which were the assessments closest to the end of the intervention, as posttest impacts. These were collected 14 months before the end of treatment, on average.
Project Alliance #2 (Danzo et al., 2020; Fosco et al., 2013, 2016; Stormshak, 2018)	2013	0.03 (0.09)	0.09 (0.14)	488	32	132	x	x		Prevent adolescent problem behaviors through a universal ecological intervention for parents with multiple levels including light-touch family resource center, parent-focused intervention that appraises child-level risks and advises on appropriate responsive parenting behaviors, as well as referrals for additional intervention.	We estimated all impacts from correlation matrices reported in papers that either used complex SEM models or were not specifically focused on identifying intervention impacts, hence differences in the statistical significance and magnitude of author-reported effects versus our effects.

Study ~ Treatment	Start year	Post ES (SE)	Adult ES (SE)	Avg *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies*)
Staying Connected ~ Parent and Adolescent Administered (Haggerty et al., 2007, 2015)	Early 2000s *	-0.13 (0.19)	0.00 (0.19)	224	2	164	x	x		Prevent middle schoolers engagement in risky behaviors through program for parents and children to strengthen family protective factors (e.g., communication, bonding, child engagement, reduced conflict) so that the family system promotes positive socialization (delivered in person).	In some cases we reach different conclusions about program impacts than the authors because we combined across subgroups, whereas the authors often focused on moderation by race.
Staying Connected ~ Self-administered with telephone support (Haggerty et al., 2007, 2015)	Early 2000s *	-0.12 (0.19)	0.05 (0.20)	213	2	164	x	x		Prevent middle schoolers engagement in risky behaviors through program for parents and children to strengthen family protective factors (e.g., communication, bonding, child engagement, reduced conflict) so that the family system promotes positive socialization (self-administered with remote oversight).	In some cases we reached different conclusions about program impacts than the authors because we combined across subgroups, whereas the authors often reported effects by moderators including race.
Strong African American Families (SAAF) (Brody et al., 2020; Miller et al., 2014)	Late 90s-Early 2000s *	-0.07 (0.34)	0.20 (0.33)	641	8	134	x	x		Prevent substance use and conduct problems among Black students through sessions with parent and/or child to strengthen family protective factors (e.g., communication, racial socialization strategies, expectations) and children's life skills (including response strategies to racist encounters, planning skills, peer pressure resistance).	We estimated some impacts from papers that were not focused specifically on identifying intervention impacts. Posttest effects were presented separately by follow-up data availability; we combined estimates.
Average	**1991**	**0.14 (0.04)**	**0.04 (0.01)**	**916**	**26**	**85**	**72%**	**93%**	**52%**		

Notes: "Avg. *n*" reflects the average total sample size at posttest for each intervention, averaged across all coded outcomes. Duration in months reflects duration from intervention start to end (e.g., an intervention that operated from the beginning of first grade until the end of second grade would be 20 months). The "Parents" "Soc-emo" and "Cog" columns indicate whether each of these factors were targeted by the treatment. "Soc-emo" refers to social-emotional skills. "Cog" refers to cognitive skills. *For some interventions, it was not possible to deduce the exact year that the intervention started; the years that are indicated are our best guess of when the intervention started. The posttest and adult effect sizes and standard errors presented for each intervention are unweighted averages for outcomes measured at the respective assessment wave. The average effect sizes presented at the bottom of the table are from Table S5, Panel B.

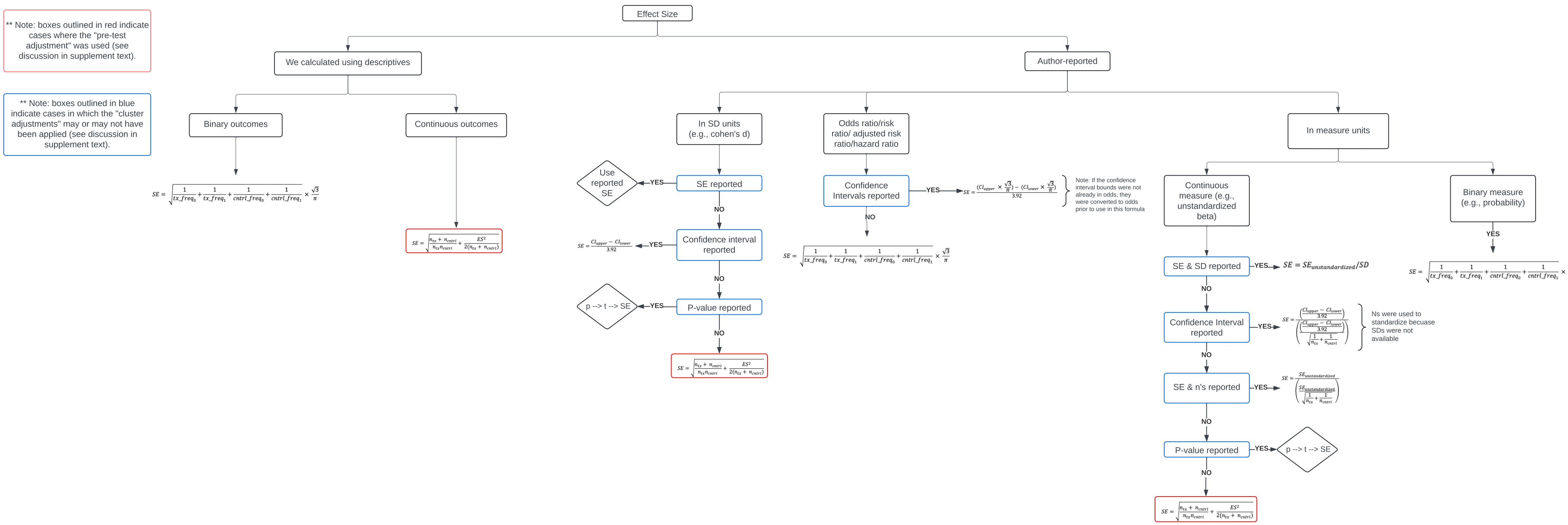
Figure S1

Effect Size Calculation Decision Flow



Notes: This decision flow chart must be viewed on a computer at high magnification. Please contact the authors for a readable print version.

Figure S2
Alternative Standard Errors Calculation Decision Flow



Notes: This decision flow chart must be viewed on a computer at high magnification. Please contact the authors for a readable print version.

Table S4
Descriptive Information on Posttest and Adult Impacts

	Mean (SE) (1)	95% Prediction Interval (2)	Studies (3)	Interventions (4)	Outcomes (5)
Posttest outcomes	**0.15 (0.05)***	**[-0.35, 0.65]**	**25**	**29**	**179**
Cognitive	0.24 (0.07)**	[-0.36, 0.83]	16	18	47
Social-emotional	0.08 (0.04)	[-0.26, 0.43]	17	21	132
Adult Outcomes	**0.11 (0.03)*****	**[-0.13, 0.35]**	**25**	**29**	**497**
Substance use	0.11 (0.04)*	[-0.18, 0.41]	13	16	137
Education	0.14 (0.04)**	[-0.18, 0.46]	16	18	112
Psychological	0.09 (0.03)*	[-0.12, 0.29]	15	17	72
Other	0.04 (0.03)	[-0.19, 0.26]	16	18	48
Employment	0.18 (0.08)	[-0.27, 0.63]	8	9	43
Cognitive	0.28 (0.14)	[-0.68, 1.23]	7	7	30
Crime	0.18 (0.07)	[-0.26, 0.61]	5	5	28
Health	0.10 (0.09)	[-0.44, 0.65]	8	9	21
Social services	0.14 (0.13)	[-0.67, 0.96]	4	4	6

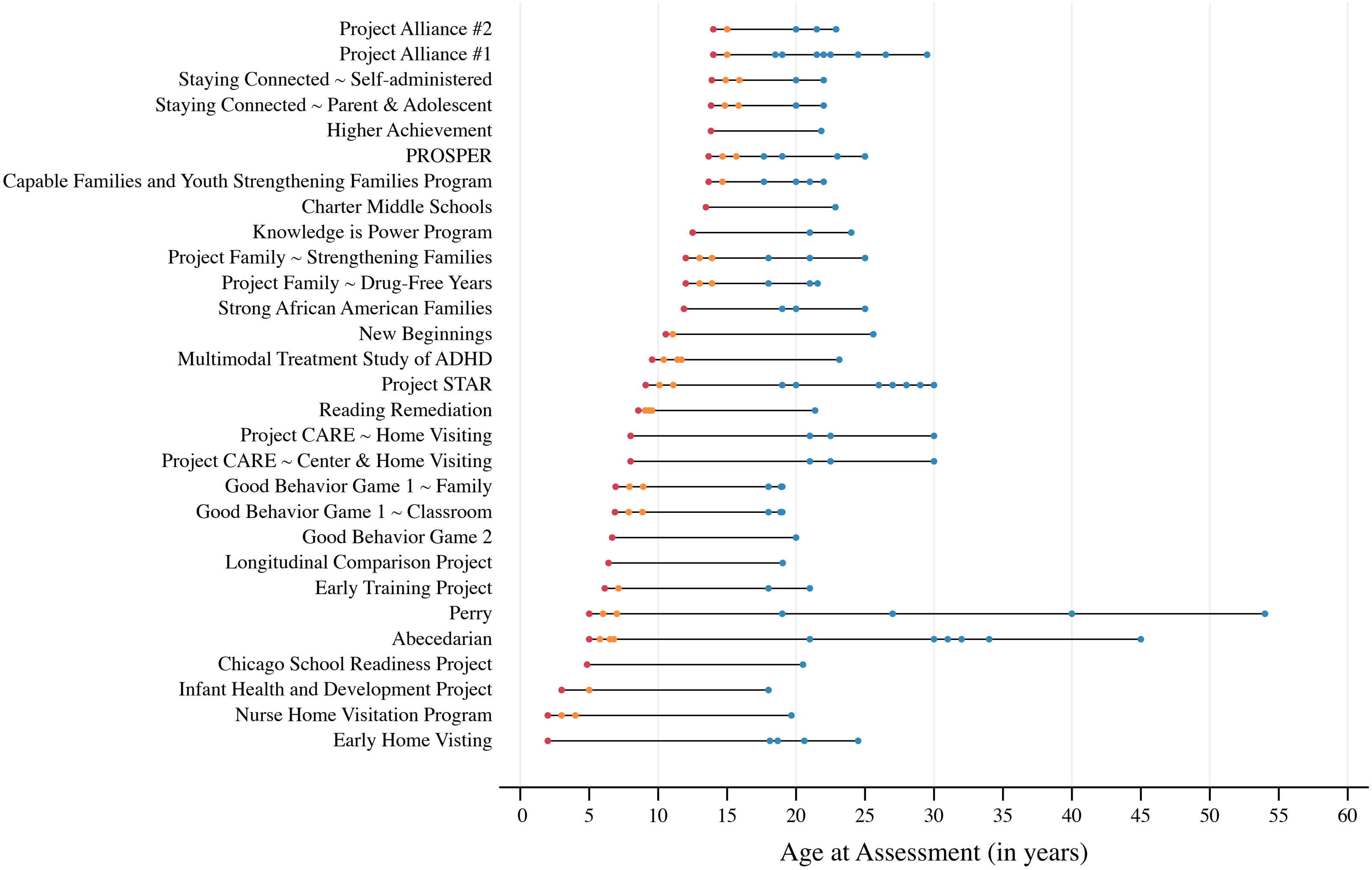
Notes: This table provides information on the outcomes that compose the posttest and adult impacts in our sample. All meta-analytic averages were estimated using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level).
* $p < .05$, ** $p < .01$, *** $p < .001$

Table S5
Average Intervention Effects Across Assessment Waves

	Posttest (1)	6-12mo follow-up (2)	1-2yr follow-up (3)	Adult follow-up (4)
Panel A: All Outcomes				
Mean (SE)	0.15 (0.05)**	0.10 (0.05)*	0.13 (0.05)*	0.11 (0.03)***
95% Prediction interval	[-0.35, 0.65]	[-0.24, 0.45]	[-0.18, 0.44]	[-0.13, 0.35]
τ_{study}	0.24	0.15	0.13	0.11
I^2	72	51	34	69
Obs (study/int/outcomes)	25 / 29 / 179	15 / 18 / 94	10 / 13 / 67	25 / 29 / 497
Panel B: Intervention-Level Averages for All Outcomes				
Mean (SE)	0.14 (0.04)**	0.07 (0.03)	0.06 (0.01)	0.04 (0.01)*
95% Prediction interval	[-0.18, 0.46]	[-0.05, 0.19]	[-0.02, 0.14]	[0.00, 0.09]
τ_{study}	0.15	0.04	0.00	0.00
I^2	67	37	0	0
Obs (study/int/outcomes)	25 / 29 / 29	15 / 18 / 18	10 / 13 / 13	25 / 29 / 29
Panel C: Intervention-Level Averages for Aligned Outcomes				
Mean (SE)	0.20 (0.11)	0.15 (0.06)*	0.08 (0.07)	0.10 (0.06)
95% Prediction interval	[-0.58, 0.99]	[-0.29, 0.58]	[-0.30, 0.46]	[-0.13, 0.32]
τ_{study}	0.33	0.18	0.11	0.07
I^2	70	48	29	0
Obs (study/int/outcomes)	10 / 13 / 57	10 / 13 / 57	5 / 8 / 24	10 / 13 / 13

Notes: All meta-analytic averages were estimated using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). Negative I^2 statistics were rounded to zero. “Aligned groups” were cases where the same intervention assessed the same construct using the same measure at posttest, 6- to 12-month follow-up, and potentially 1- to 2-year follow-up.
* $p < .05$, ** $p < .01$, *** $p < .001$

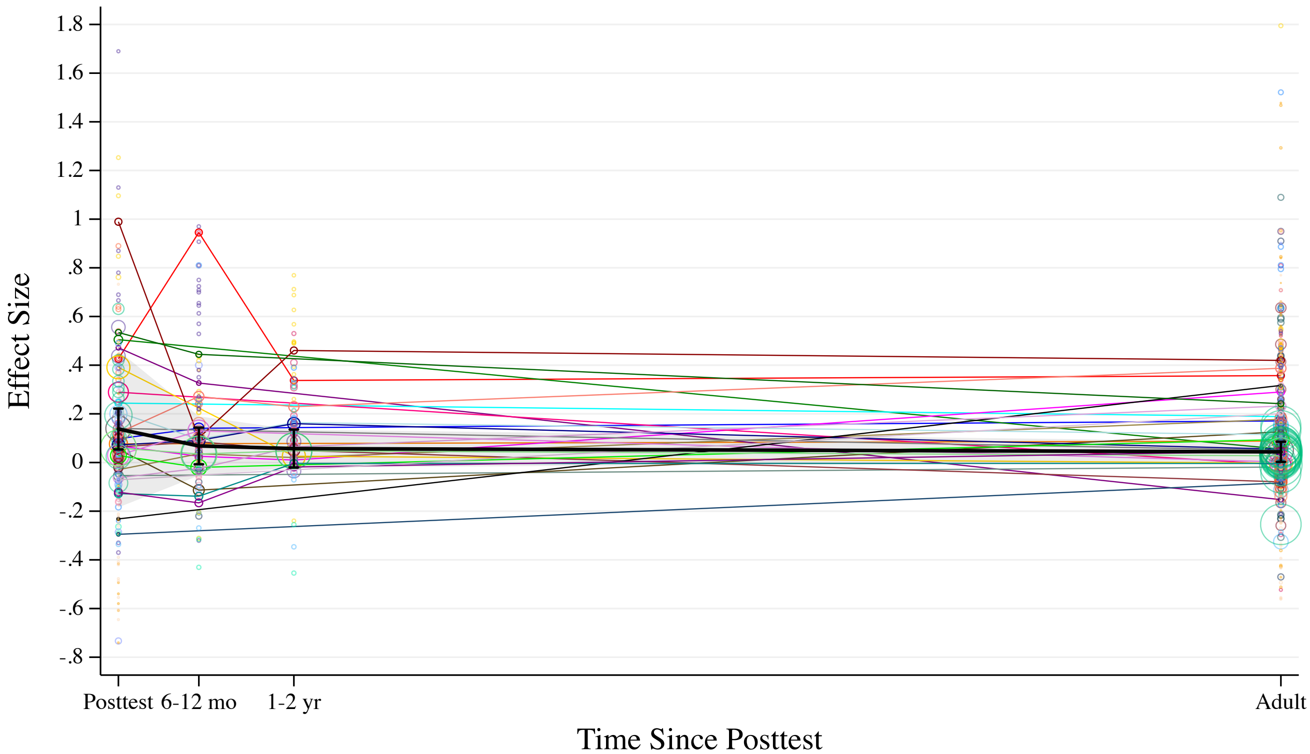
Figure S3
Intervention-Level Assessment Timelines



Notes: Each line portrays the assessments collected for each intervention in the sample. Red coordinates reflect posttest assessments. Orange coordinates reflect medium-term follow-up assessments occurring 6 to 12 months after intervention end. Blue coordinates reflect adult follow-up assessments collected as early as when participants graduated high school/turned 18 years old.

Figure S4

Trajectories of Average Effects for Each Intervention with Outcome Level Effects Displayed



Notes: Each line reflects intervention-level averages across time, with coordinates weighted by the posttest inverse sampling variances. The coordinates that are not connected to the lines reflect the individual impacts that contributed to the intervention-level averages with weighting by the concurrent inverse sampling variance. The black line depicts the meta-analytic averages at each assessment wave, estimated in R using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). The black error bars reflect 95% cluster-robust confidence intervals. The gray shaded region reflects the 95% prediction intervals. For visualization purposes, four effects that were less than -.80 *SD* were removed from this figure.

Table S6
Conditional Persistence Rates for Aligned Outcomes

	6- to 12-month follow-up (1)	1- to 2-year follow-up (2)
Intercept	0.05 (0.03)	0.04 (0.06)
Posttest	0.51 (0.13)*	0.26 (0.08)
τ_{study}	0.07	0.06
I^2	0	16
Obs (study/int/outcomes)	10 / 13 / 57	5 / 8 / 24

Notes: This table presents the association between posttest and medium-term follow-up impacts at 6- to 12-months and 1- to 2-year follow-up. Only "aligned groups" in which the same construct was assessed using the same assessment over time were included. Both models were executed using follow-up inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level).
* $p < .05$, ** $p < .01$, *** $p < .001$

Table S7
Adult Follow-Up Impacts Predicted by Posttest and Medium-term Follow-Up Impacts

				Has 6- to 12-month data			Has 1- to 2-year data		
	All outcomes (1)	All outcomes & Int RE (2)	Aligned outcomes (3)	All outcomes (4)	All outcomes (5)	Aligned outcomes (6)	All outcomes (7)	All outcomes (8)	Aligned outcomes (9)
Intercept	0.03 (0.01)	0.03 (0.01)	0.08 (0.06)	0.07 (0.04)	0.06 (0.04)	0.10 (0.06)	0.01 (0.04)	0.07 (0.05)	0.10 (0.09)
Posttest	0.13 (0.12)	0.13 (0.12)	0.14 (0.21)		0.34 (0.07)*			0.24 (0.17)	
6- to 12-month follow-up				0.27 (0.09)		0.00 (0.32)			
1- to 2-year follow-up							0.94 (0.22)*		0.71 (0.34)
τ	0.00	0.00	0.08	0.06	0.06	0.08	0.04	0.09	0.13
I^2	0	0	0	0	0	0	0	16	0
Obs (study/int)	25 / 29	25 / 29	10 / 13	15 / 18	15 / 18	10 / 13	10 / 13	10 / 13	5 / 8

Notes: This table presents estimates from regression models predicting adult follow-up impacts (intervention-level averages) using posttest and medium-term follow-up impacts. All models except for that presented in Column 2 were executed using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). Column 2 indicates the associations among posttest and adult follow-up impacts in the full sample using an intervention-level random effect (as opposed to a study-level random effect), inverse variance weighting, and cluster-robust standard errors (at the study level). Columns 1, 4, and 5 indicate estimates from models predicting intervention-level-average adult impacts by intervention-level-average posttest, 6- to 12-month follow-up, and 1- to 2-year follow-up impacts, respectively. Columns 5 and 8 report predictions from posttest to adult follow-up using only data from the interventions that had 6- to 12-month or 1- to 2-year data, respectively. Columns 3, 6, and 9 report the associations among posttest and medium-term follow-up impacts and adult impacts using only "aligned groups" in which the outcomes and instruments were consistent across posttest and medium-term follow-up waves. Negative I^2 statistics were rounded to zero.

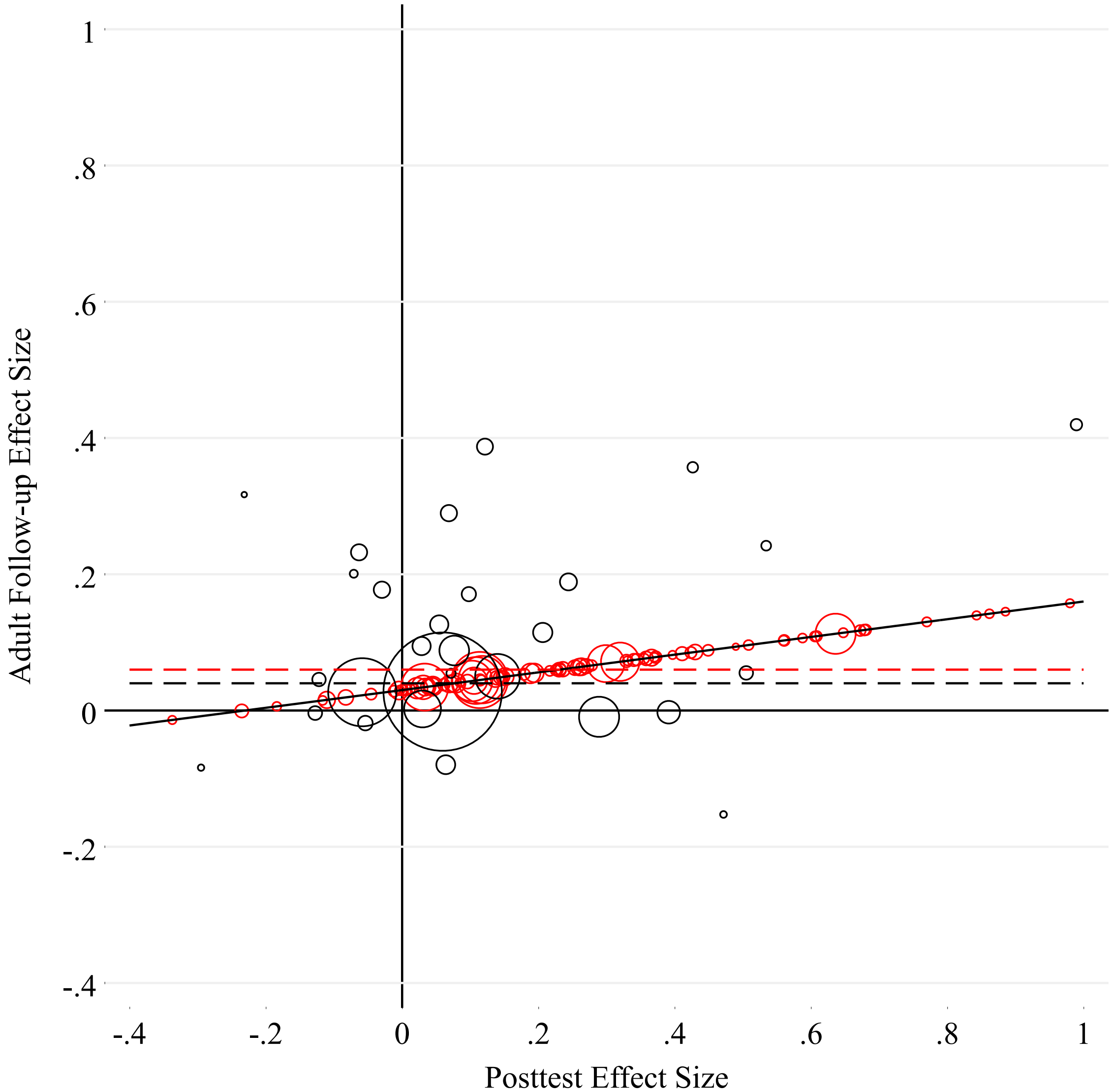
* $p < .05$, ** $p < .01$, *** $p < .001$

Table S8
Adult Follow-Up Impacts Predicted by Posttest Impacts with Each Study Dropped

Dropped Study	Slope	Intercept
Abecedarian	0.11 (0.12)	0.04 (0.01)
Capable Families and Youth Strengthening Families Program	0.13 (0.12)	0.03 (0.01)
Chicago School Readiness Project	0.14 (0.13)	0.03 (0.01)
Early Home Visiting	0.13 (0.11)	0.03 (0.01)
Early Training Project	0.14 (0.12)	0.03 (0.01)
Good Behavior Game 2	0.12 (0.12)	0.03 (0.01)
Good Behavior Game	0.14 (0.11)	0.03 (0.01)
Higher Achievement	0.21 (0.10)	0.03 (0.02)
Infant Health and Development Program	0.17 (0.13)	0.03 (0.01)
Knowledge is Power Program	0.13 (0.12)	0.03 (0.01)
Longitudinal Comparison Project	0.13 (0.12)	0.04 (0.01)
Multimodal Treatment Study of ADHD	0.13 (0.12)	0.04 (0.01)
New Beginnings	0.13 (0.11)	0.03 (0.01)
Nurse Home Visitation program	0.12 (0.12)	0.03 (0.01)
PROSPER	0.13 (0.12)	0.03 (0.01)
Perry	0.04 (0.10)	0.04 (0.01)
Project Alliance #1	0.13 (0.12)	0.04 (0.02)
Project Alliance #2	0.13 (0.11)	0.03 (0.01)
Project CARE	0.13 (0.12)	0.03 (0.01)
Project Family	0.14 (0.11)	0.03 (0.01)
Project Star	0.12 (0.12)	0.05 (0.02)
Reading Remediation	0.12 (0.12)	0.04 (0.01)
Staying Connected	0.13 (0.12)	0.03 (0.01)
Strong African American Families	0.13 (0.12)	0.03 (0.01)
Charter Middle Schools	0.13 (0.14)	0.03 (0.02)

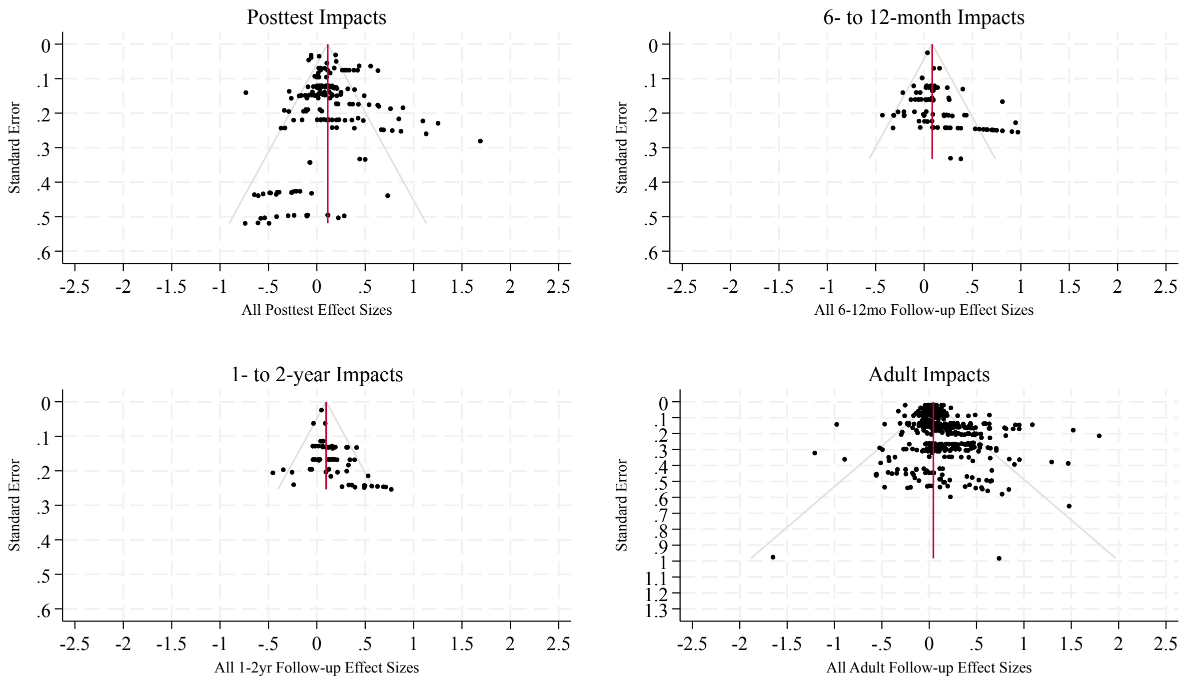
Notes: This table presents estimates from regression models predicting adult follow-up impacts using posttest impacts, both at the intervention-average level. In each model, a different study was dropped. All models were estimated using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level).
* $p < .05$, ** $p < .01$, *** $p < .001$

Figure S5
Using Our Estimates to Forecast the Long-Run Effects of Other Interventions



Notes: The black coordinates represent the average posttest and adult follow-up impact for each intervention in our sample (n = 29 intervention groups). The black solid line reflects the association between average posttest and adult follow-up effects in our sample. The black dashed line indicates the meta-analytic average of adult impacts (.04 *SD*). The red coordinates reflect intervention-level posttest averages from the MERF sample for which adult follow-up impacts were not measured and/or reported (n = 95 intervention groups). As such, the corresponding adult follow-up effect sizes reflect imputed values, estimated using the regression-based intercept and slope computed with our data. The red dashed line indicates the meta-analytic average of these imputed values (.06 *SD*; posttest standard errors were used as weights). All coordinates are weighted by the inverse sampling variances. (For observed adult impacts from our sample, the adult impact variances are used. For imputed adult impacts from the MERF sample, weighting relies on the variances of the posttest effects.) To maintain the same axes as other figures, this figure drops 7 interventions from the MERF sample that had average posttest impacts that were greater than 1.

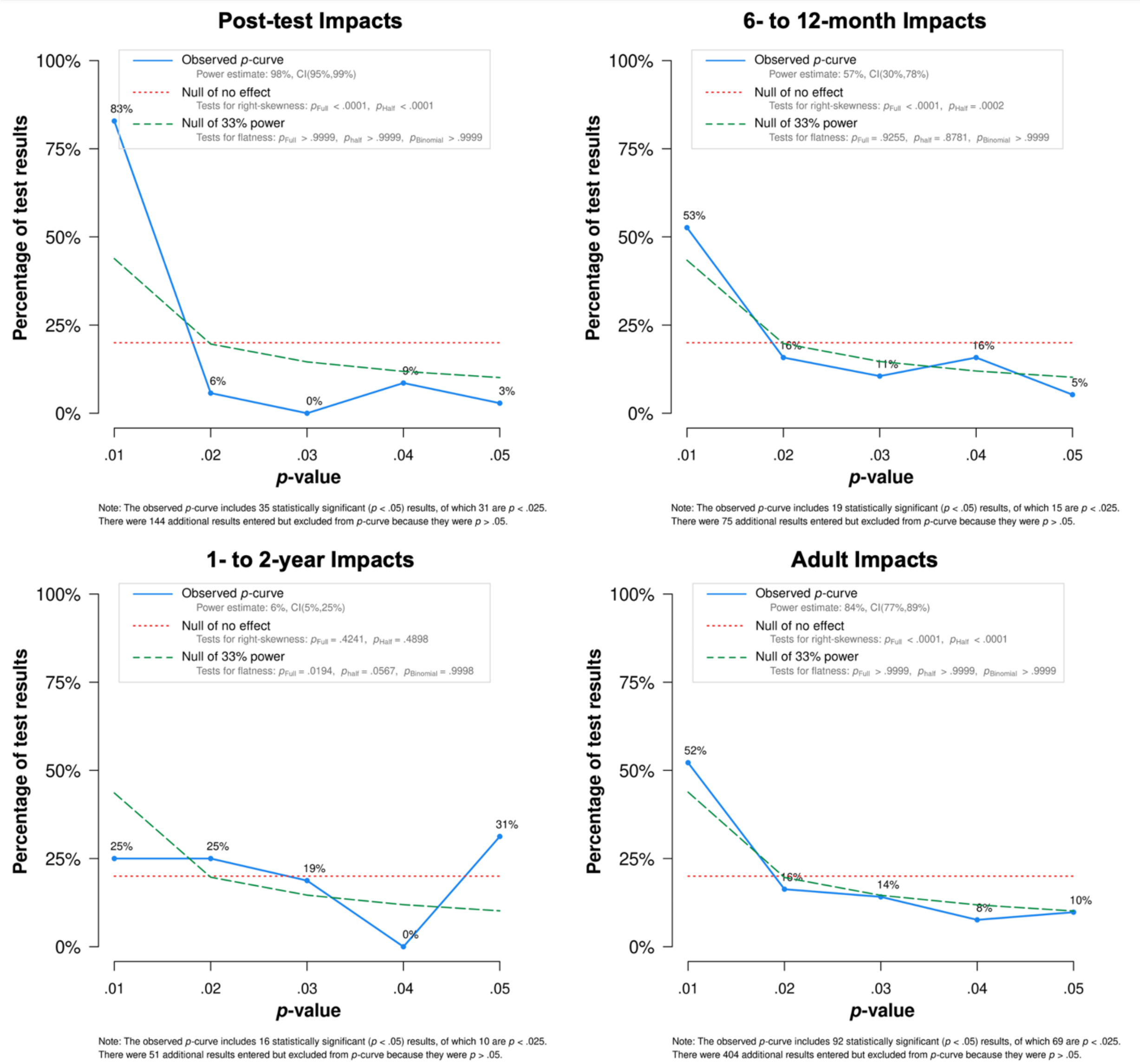
Figure S6
Funnel Plots for Posttest, Medium-Term Follow-Up, and Adult Follow-Up Impacts



Notes: Each coordinate reflects one effect size for one outcome. Gray lines represent 95% confidence intervals. Note that the standard error scale is much larger than that for posttest and medium-term follow-up impacts.

Figure S7

P-curves for Posttest, Medium-Term Follow-Up, and Adult Follow-Up Impacts



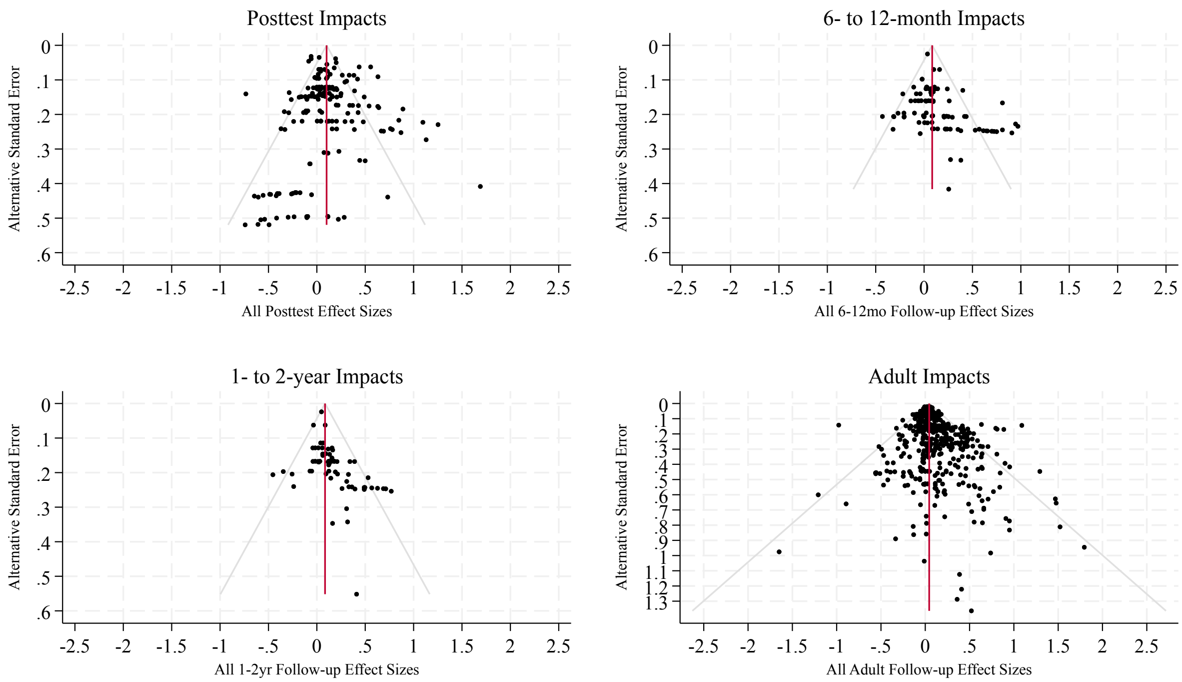
Notes: Each figure presents the *p*-curve for *p*-values of the outcomes contributing effects at each assessment wave. Figures were created on p-curve.com (Simonsohn, Nelson, & Simmons, 2015).

Table S9
Selection Models

Assessment Wave	Original Est [95% CI] (1)	Adjusted Est [95% CI] (2)	Adjusted Est / Original Est (3)
Panel A: Stanley & Doucouliagos (2014) PET Model			
Posttest	0.15 [0.05, 0.26]	0.09 [-0.06, 0.24]	60%
6- to 12-mo Follow-up	0.10 [0.01, 0.20]	0.02 [-0.06, 0.10]	20%
1- to 2-year Follow-up	0.13 [0.02, 0.23]	0.03 [-0.03, 0.09]	23%
Adult Follow-up	0.11 [0.06, 0.17]	0.01 [-0.03, 0.05]	9%
Panel B: Vevea & Woods (2005) Selection Model			
Posttest	0.15 [0.05, 0.26]	0.12 [0.08, 0.16]	80%
6- to 12-mo Follow-up	0.10 [0.01, 0.20]	0.12 [0.07, 0.17]	120%
1- to 2-year Follow-up	0.13 [0.02, 0.23]	0.13 [0.09, 0.17]	100%
Adult Follow-up	0.11 [0.06, 0.17]	0.11 [0.09, 0.13]	100%
Panel C: Pustejovsky, Joshi, & Citkowicz (2025) Selection Model (2 steps)			
Posttest	0.15 [0.05, 0.26]	0.22 [0.06, 0.38]	147%
6- to 12-mo Follow-up	0.10 [0.01, 0.20]	0.02 [-0.03, 0.22]	20%
1- to 2-year Follow-up	0.13 [0.02, 0.23]	0.03 [-0.01, 0.07]	23%
Adult Follow-up	0.11 [0.06, 0.17]	0.15 [0.03, 0.29]	136%
Panel D: Pustejovsky, Joshi, & Citkowicz (2025) Selection Model (9 steps)			
Posttest	0.15 [0.05, 0.26]	0.28 [-0.45, 0.48]	187%
6- to 12-mo Follow-up	0.10 [0.01, 0.20]	0.17 [-0.04, 0.46]	170%
1- to 2-year Follow-up	0.13 [0.02, 0.23]	-0.00 [-0.29, 0.12]	0%
Adult Follow-up	0.11 [0.06, 0.17]	0.06 [-0.18, 0.29]	55%

Note. Panel A reports the original effect sizes and adjusted effect sizes from the PET model which controls for the effect sizes' accompanying standard error (Stanley & Doucouliagos, 2014). As such, the adjusted estimates reflect anticipated intervention impacts when effect sizes are perfectly precise (i.e., zero). Panel B reports original and adjusted effect sizes estimated using a weight-function model with cut-points and weights as detailed in Vevea and Woods (2005). Weights were set as follows to reflect patterns of selective reporting if p-values dictated publishing: effects associated with a $p < 0.01$ were set to have a weight of 1 (an assumption that 100% of effects of this statistical significance are published if selection biases are at play), $p < .05$ is set to .9, $p < .50$ is set to .70, and $p < 1$ is set to .50. Panel C reports original and adjusted estimates using an alternative weight function that accounts for dependencies in the data developed by Pustejovsky, Joshi, and Citkowicz (2025). The model estimated the adjusted effect as well as the likelihood that findings with $p > .025$ to $p < .50$ and $p > .50$ to $p < 1$ would be reported relative to $p < .025$. The model was estimated with 1,000 bootstraps (bootstrap = "two-stage"). Panel D reports estimates using a 9-step version of the Pustejovsky et al. selection model with steps at .025, .05, .10, .20, .30, .40, .50, .70, and .90.

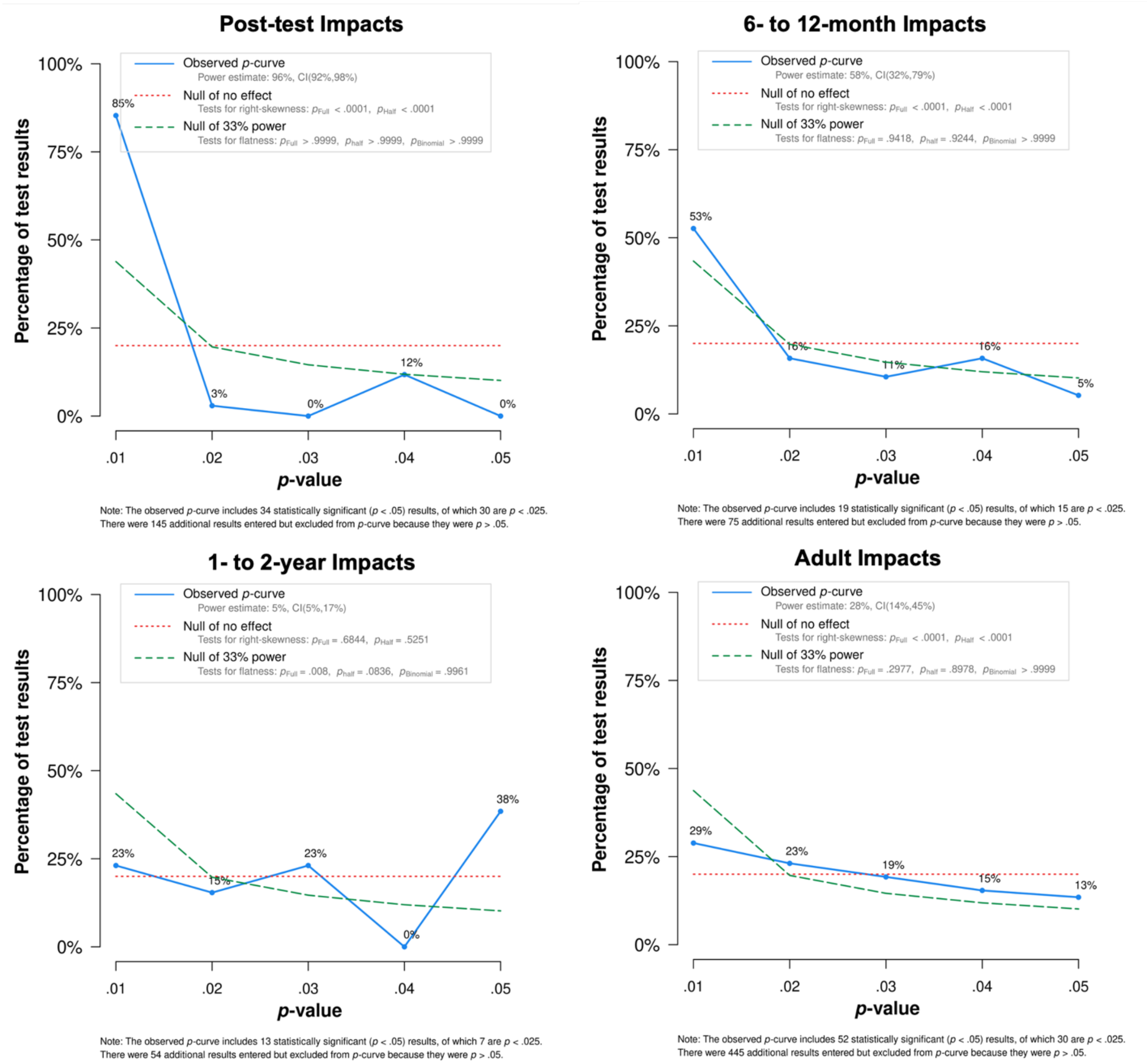
Figure S8
Funnel Plots Using Alternative Standard Errors



Notes: Each coordinate reflects one effect size for one outcome. Gray lines represent 95% confidence intervals. Note that the standard error scale is much larger than that for posttest and medium-term follow-up impacts. Standard errors reflect alternative standard errors computed through more complex methods (see Figure S2).

Figure S9

P-curves Using Alternative Standard Errors



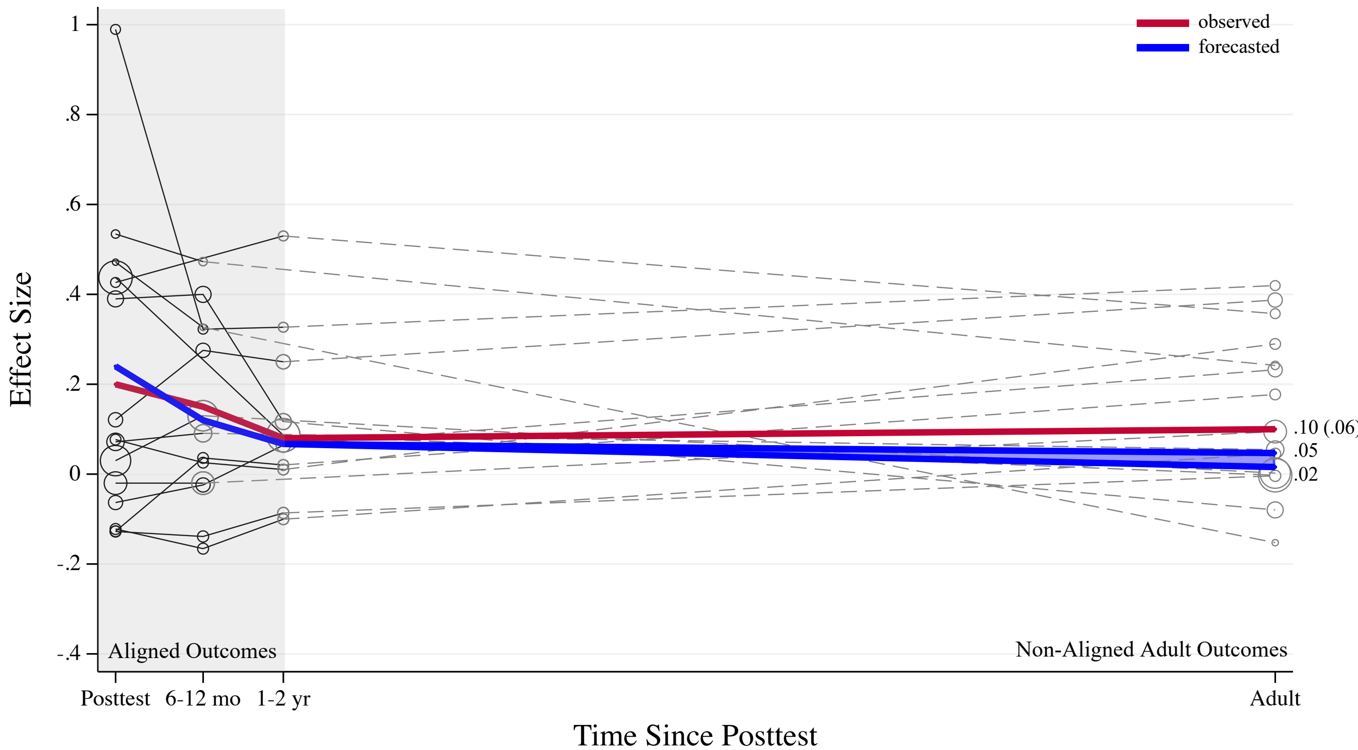
Notes: Each figure presents the *p*-curve for *p*-values of the outcomes contributing effects at each assessment wave. *P*-values were estimated using the alternative standard errors computed through more complex techniques. Figures were created on p-curve.com (Simonsohn, Nelson, & Simmons, 2015). We used alternative standard errors computed through more complex methods (see Figure S2).

Table S10
Primary Results Using Alternative Weighting Approaches and the Correlated-and-Hierarchical Model

	Averages				Predicting adult impacts (5)
	Posttest (1)	6- to 12-month follow-up (2)	1- to 2-year follow-up (3)	Adult (4)	
Panel A: No Weights					
Intercept	0.13 (0.06)*	0.13 (0.06)	0.12 (0.04)*	0.11 (0.03)**	0.09 (0.03)*
Posttest					0.15 (0.12)
τ_{study}	0.15	0.04	0.00	0.00	0.00
I^2	67	37	0	0	0
Obs (study / int)	25 / 29	15 / 18	10 / 13	25 / 29	25 /29
Panel B: Alternative SE Calculations					
Intercept	0.13 (0.04)**	0.07 (0.03)	0.06 (0.01)	0.04 (0.01)*	0.03 (0.01)
Posttest					0.13 (0.13)
τ_{study}	0.14	0.04	0.00	0.00	0.00
I^2	66	37	0	0	0
Obs (study / int)	25 / 29	15 / 18	10 / 13	25 / 29	25 /29
Panel C: Sample Size Weighting					
Intercept	0.16 (0.05)**	0.14 (0.06)*	0.11 (0.03)*	0.10 (0.02)***	0.07 (0.02)**
Posttest					0.21 (0.11)
τ_{study}	0.24	0.23	0.09	0.10	0.10
I^2	94	88	84	82	82
Obs (study / int)	25 / 29	15 / 18	10 / 13	25 / 29	25 /29
Panel D: Correlated-and-Hierarchical Effects Model					
Intercept	0.14 (0.04)**	0.09 (0.03)*	0.10 (0.03)*	0.06 (0.02)**	0.03 (0.01)
Posttest					0.16 (0.11)
τ_b / τ_w	0.14 / 0.17	0.00 / 0.17	0.00 / 0.11	0.00 / 0.21	0.00 / 0.00
I^2	78	66	52	85	0
Obs (study / int / outcome)	25 / 29 / 179	15 / 18 / 94	10 / 13 / 67	25 / 29 / 497	25 / 29
Panel E: Avg. Posttest Linked with Each Adult Outcome (Correlated-and-Hierarchical Effects Model)					
Intercept	0.15 (0.05)**			0.06 (0.02)**	0.03 (0.01)*
Posttest					0.32 (0.09)*
τ_b / τ_w	0.19 / 0.00			0.00 / 0.21	0.00 / 0.20
I^2	0			85	85
Obs (study / int / outcome)	25 / 29 / 497			25 / 29 / 497	25 / 29 / 497

Notes: This table presents the meta-analytic averages at posttest, medium-term follow-up, and adult impacts and the association between posttest and adult impacts using three different weighting approaches. Models were estimated using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). Negative I^2 statistics were rounded to zero. Panel A presents estimates computed without weighting. Panel B reports estimates from models using alternative standard errors that were computed with more intensive calculations (see Figure S2 for details). Panel C presents estimates from a model in which sample-size weights were used (i.e., for columns 1, 2, and 3 concurrent intervention-level average sample sizes were used, in column 4 the intervention-level average of sample sizes for adult impacts was used). Panel D presents estimates from the CHE model (Pustejovsky & Tipton, 2022). Columns 1, 2, 3, and 4 were estimated with outcome level data and column 4 was estimated with intervention-level averages. We estimated between (τ_b) and within (τ_w) study variances. Effects from the same study were assumed to correlate at r = .60. Panel E presents estimates from the CHE model (with r = .60) in which intervention-level average posttest impacts were matched with outcome-level adult impacts. As such, there were only 29 unique intervention-level post-test values linked with 497 outcome-level adult effects.
* $p < .05$, ** $p < .01$, *** $p < .001$

Figure S10
LINT Simulation and Trajectories of Effects for Interventions with Aligned Medium-term Impacts



Notes: Each line reflects one intervention that measured the same measure and construct at posttest and at least one medium-term follow-up assessment wave. The posttest, 6- to 12-month and 1- to 2-year follow-up coordinates (shaded in gray) reflect the intervention-level average of the intervention impacts that were estimated consistently across assessment waves. The adult coordinate reflects the average of the intervention's impact on all outcomes measured in adulthood. Coordinates are weighted by the posttest inverse sampling variances. The red line depicts the meta-analytic averages at each assessment wave, estimated using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). The blue line depicts the forecasted meta-analytic averages at each assessment wave (see text for more details). See Table S5, Panel C for estimates.

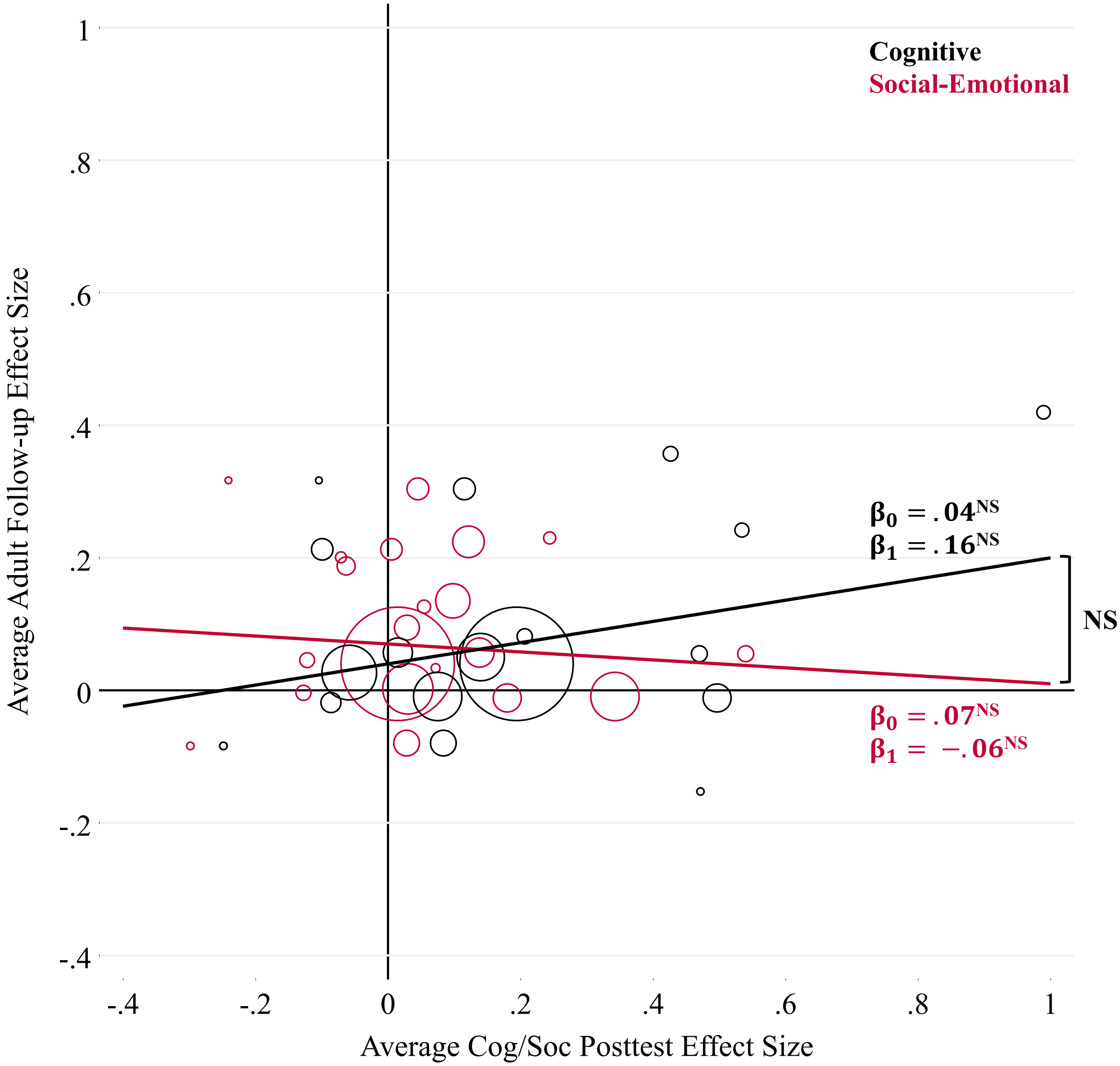
* $p < .05$, ** $p < .01$, *** $p < .001$

Table S11
Outcome-Type Differences in Meta-Analytic Averages and the Associations Between Posttest and Adult Impacts using Intervention-level Averages

	Averages by Outcome Types		Predicting Adult Impacts	
	All (1)	Outcome type interaction (2)	All (3)	Outcome type interaction (4)
Panel A: Posttest impacts (social-emotional versus cognitive)				
Intercept	0.14 (0.04)**	0.09 (0.05)	0.04 (0.01)*	0.07 (0.03)
Cog (vs. Soc) outcome		0.10 (0.07)		-0.03 (0.02)
Posttest			0.05 (0.07)	-0.06 (0.13)
Posttest x Cog				0.22 (0.18)
τ_{study}	0.14	0.14	0.02	0.04
I^2	71	71	0	0
Obs (study/int/outcomes)	25 / 29 / 39	25 / 29 / 39	25 / 29 / 39	25 / 29 / 39
Panel B: Adult impacts by outcome category				
Intercept	0.08 (0.02)**	0.06 (0.04)	0.05 (0.02)*	0.02 (0.04)
Posttest			0.23 (0.12)	0.29 (0.13)
Educational outcome		0.06 (0.03)		0.08 (0.03)
Psychological outcome		0.02 (0.06)		0.01 (0.07)
Substance use outcome		0.02 (0.06)		0.01 (0.06)
Posttest x Educational				-0.20 (0.11)
Posttest x Psychological				0.00 (0.11)
Posttest x Substance				0.12 (0.18)
τ_{study}	0.07	0.08	0.07	0.08
I^2	17	14	10	8
Obs (study/int/outcomes)	25 / 29 / 71	25 / 29 / 71	25 / 29 / 71	25 / 29 / 71

Notes: All estimates reported in the table were estimated using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). Panel A examines the association between social-emotional versus cognitive posttest impacts and adult impacts. Social-emotional and cognitive intervention-level averages were computed using all of the available cognitive and social-emotional measures and then matched with the intervention-level average of adult impacts. Column 1 presents the meta-analytic average of social-emotional and cognitive posttests. Column 2 indicates how posttest impacts varied by outcome type. Columns 3 and 4 show the association between posttest and adult impacts and how this varies by outcome type. Panel B examines how the association between posttest and adult impacts varies by adult outcome type. Columns 1 and 2 indicate adult averages whereas Columns 3 and 4 show the regression model with and without outcome category interactions. Negative I^2 statistics were rounded to zero.
* $p < .05$, ** $p < .01$, *** $p < .001$

Figure S11
Posttest Social-Emotional and Cognitive Impacts Predicting Adult Impacts



Notes: "NS" = Statistically non-significant. This figure plots the average cognitive (black) and social-emotional (red) posttest effect size for each intervention and the corresponding intervention-level average adult impact. Coordinates were weighted by the follow-up inverse sampling variances. The lines depict the estimated intercepts (B_0) and slopes (B_1) for cognitive and social-emotional posttest impacts estimated in R using follow-up inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). See Table S11 for estimates.
* $p < .05$, ** $p < .01$, *** $p < .001$

Table S12
Outcome-Type Differences in the Association between Post-test and Adult Effects

	Educational (1)	Substance (2)	Psychological (3)
Intercept	0.11 (0.04) *	0.06 (0.05)	0.08 (0.04)
Cog (vs. Soc) outcome	-0.04 (0.03)	0.00 (0.03)	-0.09 (0.06)
Posttest	-0.11 (0.08)	0.53 (0.24)	-0.31 (0.42)
Posttest x Cog	0.28 (0.18)	-0.17 (0.27)	0.60 (0.46)
τ_{study}	0.10	0.11	0.06
I^2	35	10	0
Obs (study/int/outcomes)	16 / 18 / 27	13 / 16 / 21	14 / 16 / 23

Notes: This table reports estimates of the differential associations among intervention-level average posttest cognitive and social-emotional impacts and intervention-level average adult impacts on the following adult outcome categories: educational, psychological wellbeing, and substance-related. All estimates were estimated using follow-up inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). Negative I^2 statistics were rounded to zero.
* $p < .05$, ** $p < .01$, *** $p < .001$

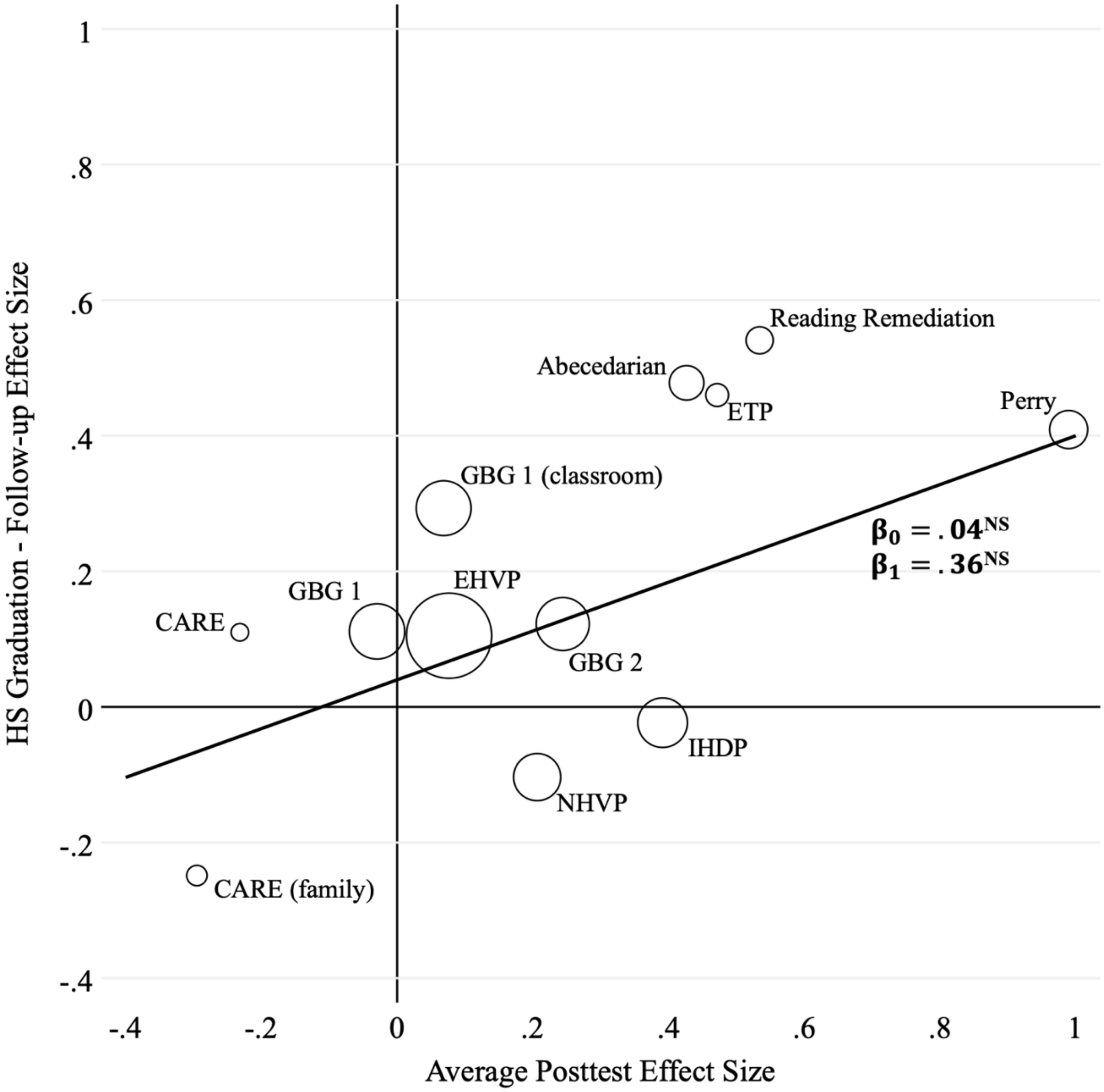
Table S13
Intervention-Level Average Posttest Impacts Predicting High School Attainment

	Posttest Average (1)	High School Attainment Average (2)	Predicting High School Attainment (3)
Intercept	0.30 (0.09)*	0.11 (0.05)	0.04 (0.07)
Posttest			0.36 (0.16)
τ_{study}	0.23	0.11	0.11
I^2	61	12	12
Obs (study/int)	10 / 12	10 / 12	10 / 12

Notes: This table presents estimates from regression models predicting adult follow-up impacts on high school attainment using intervention-level average posttest effects. Twelve interventions contributed high school attainment impact estimates; in the case that an intervention measured this outcome multiple times, the earliest assessment was used to increase cross-intervention harmony (for all interventions, this outcome was assessed when participants were 18 to 22 years old). Estimates were generated using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). Negative I^2 statistics were rounded to zero.

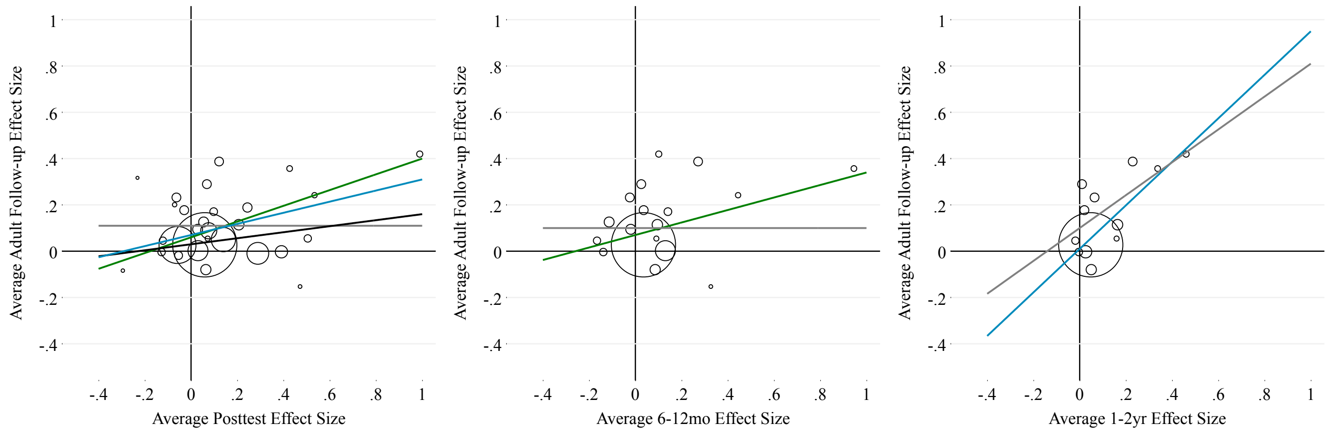
* $p < .05$, ** $p < .01$, *** $p < .001$

Figure S12
Posttest Impacts Predicting High School Graduation



Notes: "NS" = Statistically non-significant. This figure plots the average posttest effect size and the follow-up effect on high school graduation. For interventions that reported impacts on this outcome at multiple adult assessment waves, we used the earliest reporting to further increase comparability across studies. Coordinates are weighted by the follow-up inverse sampling variances. The line depicts the estimated intercept (B_0) and slope (B_1) estimated in R using follow-up inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). To reduce clutter, for studies with two intervention groups, the coordinate label without parentheses reflects the group that is not otherwise listed on the figure. See Table S3 for full intervention names associated with the label abbreviations and Table S13 for estimates.

Figure S13
Posttest and Medium-Term Follow-Up Impacts Predicting Adult Impacts



Notes: Figure A plots the association between the intervention-level average posttest impact and adult impact. Figures B and C depict the association between intervention-level average effects at medium-term follow-up assessments (6- to 12-month follow-up and 1- to 2-year follow-up) and in adulthood. Each coordinate represents one study and was weighted by the inverse sampling variances. The black line depicts the estimated intercept (B_0) and slope (B_1) using intervention-level averages at posttest and follow-up. The gray lines depict the estimated intercept and slope using aligned groups (i.e., the same construct and measure across time and models). The green lines indicate the estimated intercept and slope using intervention-level averages within the sample of studies that had 6- to 12-month follow-up data. The blue lines indicate the estimated intercept and slope using intervention-level averages within the sample of studies that had 1- to 2-year follow-up data. All models were executed in R using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). See Table S7 for estimates.

Table S14
Intervention-Level Average Posttest Impacts Predicting Intervention-Level Average Adult Impacts at Each Adult Follow-up Assessment Wave

	Averages		Predicting Adult Impacts		
	Posttest (1)	Adult (2)	Baseline (3)	Main Effect (4)	Interaction (5)
Intercept	0.15 (0.05)**	0.12 (0.03)***	0.09 (0.03)*	0.07 (0.03)*	0.08 (0.03)*
Posttest			0.25 (0.12)	0.33 (0.16)	0.26 (0.15)
Adult Age				-0.01 (0.00)	-0.01 (0.00)
Posttest x Age					0.01 (0.01)
τ_{study}	0.23	0.12	0.11	0.12	0.11
I^2	54	79	78	75	75
Obs (study/int)	25 / 29 / 79	25 / 29 / 79	25 / 29 / 79	25 / 29 / 79	25 / 29 / 79

Notes: This table presents estimates from regression models examining the associations among intervention-level average posttest impacts and intervention-level average adult impacts, aggregated at each adult follow-up assessment wave. "Adult age" is a continuous indicator for adult age at follow-up, centered at the mean of the distribution. Estimates were generated using inverse sampling variance weighting (by follow-up effects in Column 3), a random effect for study, and cluster-robust standard errors (with clustering at the study level). Negative I^2 statistics were rounded to zero.

* $p < .05$, ** $p < .01$, *** $p < .001$

Table S15
Examining Assessment-Timing-Related Concerns in the Primary Models

	Adult Averages			Predicting Adult Impacts		
	Months (1)	Assess Age (2)	ECE (3)	Months (4)	Assess Age (5)	ECE (6)
Intercept	0.03 (0.05)	0.10 (0.11)	0.04 (0.02)	0.03 (0.05)	0.09 (0.11)	0.04 (0.02)
Posttest				0.13 (0.14)	0.13 (0.13)	0.13 (0.13)
Months from Posttest at Assessment	0.00 (0.00)			0.00 (0.00)		
Adult Age		0.00 (0.00)			0.00 (0.00)	
ECE vs. Non-ECE Intervention			0.01 (0.04)			0.00 (0.04)
τ_{study}	0.03	0.00	0.02	0.02	0.00	0.01
I^2	0	0	0	0	0	0
Obs (study / int)	25 / 29	25 / 29	25 / 29	25 / 29	25 / 29	25 / 29

Notes: "ECE" = Early Childhood Education. This table presents the meta-analytic average of adult impacts and the association between posttest and medium-term follow-up impacts with controls for three age-/time-related variables. Models were estimated using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). The first variable that we controlled for was the months that elapsed between posttest and adult assessment. The second variable that we controlled for was adult age at adult follow-up. And the third variable indicated whether the intervention was an ECE intervention (i.e., targeted children 7 years or younger) or not. Negative I^2 statistics were rounded to zero.

* $p < .05$, ** $p < .01$, *** $p < .001$

Table S16
Probing Study and Data Irregularities

	Averages				Predicting adult impacts
	Posttest (1)	6- to 12-month follow-up (2)	1- to 2-year follow-up (3)	Adult (4)	(5)
Panel A: High-Attrition Adult Outcomes Dropped					
Intercept	0.14 (0.04)**			0.05 (0.02)	0.04 (0.02)
Posttest					0.17 (0.13)
τ_{study}	0.15			0.00	0.00
I^2	63			0	0
Obs (study / int)	20 / 24			20 / 24	20 / 24
Panel B: Posttests Collected Before End of Treatment Dropped					
Intercept	0.16 (0.05)**			0.06 (0.02)	0.05 (0.03)
Posttest					0.10 (0.16)
τ_{study}	0.16			0.03	0.03
I^2	61			0	0
Obs (study / int)	21 / 25			21 / 25	21 / 25
Panel C: Alternative Domain-Level Aggregation					
Intercept				0.04 (0.02)	0.03 (0.02)
Posttest					0.14 (0.13)
τ_{study}				0.03	0.03
I^2				0	0
Obs (study / int)				25 / 29	25 / 29
Panel D: "Lifetime"/"Ever" Measures Dropped					
Intercept	0.14 (0.04)**	0.07 (0.03)	0.06 (0.02)	0.04 (0.01)*	0.04 (0.01)*
Posttest					0.11 (0.12)
τ_{study}	0.15	0.04	0.00	0.00	0.00
I^2	67	38	0	0	0
Obs (study / int)	25 / 29	15 / 18	10 / 13	25 / 29	25 / 29
Panel E: "High-Estimation" Effects Dropped					
Intercept	0.13 (0.04)**	0.06 (0.03)	0.05 (0.01)*	0.05 (0.02)	-0.01 (0.02)
Posttest					0.61 (0.19)^
τ_{study}	0.13	0.03	0.00	0.03	0.00
I^2	64	40	0	0	0
Obs (study / int)	24 / 27	14 / 17	9 / 12	19 / 22	18 / 20

Notes: This table presents the results from several models probing study and data irregularities. The meta-analytic averages at posttest, medium-term follow-up, and adult impacts and the association between posttest and adult impacts is provided, as appropriate for the specific robustness check. Models were estimated using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). Panel A presents the models with high-attrition (> 20 %) adult outcomes dropped. Panel B presents the estimates having dropped studies that did not have true end-of-treatment impacts (i.e., "posttest" was computed when the treatment was still ongoing). Panel C presents the estimates having formed the intervention-level averages for adult outcomes by averaging outcome-domain averages (versus averaging all outcomes regardless of domain). Panel D drops outcomes that were coded as being "lifetime" or "ever" measures. Panel E drops outcomes that required a great deal of computation when estimating the effect size. ^ The Perry Preschool datapoint had an outsized impact on this slope. After dropping Perry, the slope estimate was 0.37 (0.21) and the intercept estimate was 0.01 (0.03). Negative I^2 statistics were rounded to zero.
* $p < .05$, ** $p < .01$, *** $p < .001$

Table S17
Associations between Intervention-level Average Posttest Impacts and Adult Impacts Under Various Assumptions about Correlated Error

r	Term	Coef. (SE)	95% CI
0.00	Intercept	0.04 (0.03)	[-0.01, 0.11]
0.00	Slope	0.14 (0.21)	[-0.28, 0.57]
0.10	Intercept	0.05 (0.03)	[-0.01, 0.11]
0.10	Slope	0.10 (0.21)	[-0.34, 0.51]
0.20	Intercept	0.05 (0.03)	[0.00, 0.12]
0.20	Slope	0.07 (0.21)	[-0.38, 0.48]
0.30	Intercept	0.05 (0.03)	[0.00, 0.12]
0.30	Slope	0.04 (0.21)	[-0.39, 0.45]
0.40	Intercept	0.06 (0.03)	[0.00, 0.13]
0.40	Slope	0.00 (0.21)	[-0.44, 0.40]
0.50	Intercept	0.06 (0.03)	[0.01, 0.13]
0.50	Slope	-0.03 (0.21)	[-0.48, 0.36]
0.60	Intercept	0.06 (0.03)	[0.01, 0.13]
0.60	Slope	-0.06 (0.21)	[-0.52, 0.31]
0.70	Intercept	0.07 (0.03)	[0.01, 0.13]
0.70	Slope	-0.09 (0.20)	[-0.54, 0.27]
0.80	Intercept	0.07 (0.03)	[0.02, 0.14]
0.80	Slope	-0.12 (0.20)	[-0.56, 0.24]
0.90	Intercept	0.07 (0.03)	[0.02, 0.15]
0.90	Slope	-0.16 (0.20)	[-0.61, 0.20]

Notes: This table relies on the methods described in Gilbert & Himmelsbach (2026).

References

Ansari, A. (2019). The persistence of preschool effects from early childhood through adolescence. *Journal of Educational Psychology, 110*(7), 952–973. https://doi.org/10.1037/edu0000255

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

Barnum, M. (2017). *74 interview: Harvard researcher David Deming takes the long view on Head Start, integration.* The 74. https://www.the74million.org/article/74-interview-harvard-researcher-david-deming-takes-the-long-view-on-head-start-integration

Besser, N., Linberg, A., Dornheim, D., Weinert, S., Roßbach, H.-G., & Lehrl, S. (2025). Fostering toddlers' numeracy and mathematical language skills through a professional development intervention on interaction quality in toddler classrooms. *Early Childhood Research Quarterly, 72,* 44–55. https://doi.org/10.1016/j.ecresq.2025.02.004

Brooks-Gunn, J. (2003). Do you believe in magic? What we can expect from early childhood intervention programs. *Social Policy Report, 17*(1), 1–16. https://doi.org/10.1002/j.2379-3988.2003.tb00020.x

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). *$320,000 kindergarten teachers.* Opportunity Insights. https://opportunityinsights.org/wp-content/uploads/2018/03/star_summary.pdf

Dizikes, P. (2023, February 8). *Study: Preschool gives a big boost to college attendance.* MIT News. https://news.mit.edu/2023/study-preschool-gives-big-boost-college-attendance-0208

Farley, R. (2022, February 8). *Biden stretches evidence for universal pre-k.* FactCheck.org. https://www.factcheck.org/2022/02/biden-stretches-evidence-for-universal-pre-k/

Garber, K. L., Burchinal, M., Kuperman, S., Foster, T. J., Campbell, F., & Ramey, C. (2026). Piecing together the puzzle: A mixed methods study of the Abecedarian Intervention. *Early Childhood Research Quarterly, 74*(1), 149–163. https://doi.org/10.1016/j.ecresq.2025.08.009

Gilbert, J. B., & Himmelsbach, Z. (2026). *Why fadeout is (probably) worse than we think: adjusting for correlated sampling error in meta-analyses of behavioral interventions.* EdWorking Papers. https://doi.org/10.26300/87r9-qm15

Gopnik, A. (2021, May 28). *Preschool's 'sleeper' effect on later life.* The Wall Street Journal. https://www.wsj.com/us-news/education/preschools-sleeper-effect-on-later-life-11622146543

King, L. S., Guyon-Harris, K. L., Valadez, E. A., Radulescu, A., Fox, N. A., Nelson, C. A., Zeanah, C. H., & Humphreys, K. L. (2023). A comprehensive multilevel analysis of the Bucharest Early Intervention Project: Causal effects on recovery from early severe deprivation. *American Journal of Psychiatry, 180*(8), 573–583. https://doi.org/10.1176/appi.ajp.20220672

Lehrer, J. (2010, July 14). *How preschool changes the brain.* Wired. https://www.wired.com/2010/07/how-preschool-changes-the-brain

Morawska, A., Dittman, C. K., & Rusby, J. C. (2019). Promoting self-regulation in young children: The role of parenting interventions. *Clinical Child and Family Psychology Review, 22*, 43–51. https://doi.org/10.1007/s10567-019-00281-5

O'Donnell, E. (2010, October). *Raj Chetty: Good kindergarten teachers boost pupils' lifetime earnings.* Harvard Magazine. https://www.harvardmagazine.com/2010/10/kindergarten-matters

Padilla-Walker, L. M., Millett, M. A., & Memmott-Elison, M. K. (2020). Can helping others strengthen teens? Character strengths as mediators between prosocial behavior and adolescents' internalizing symptoms. *Journal of Adolescence, 79*, 70–80. https://doi.org/10.1016/j.adolescence.2020.01.001

Petrilli, M. J. (2018). *In-depth: Michael Petrilli's 5-part argument about school choice, tests & long-term outcomes.* The 74. https://www.the74million.org/article/in-depth-michael-petrillis-5-part-argument-about-school-choice-tests-long-term-outcomes/

Robson, D. A., Alen, M. S., & Howard, S. J. (2020). Self-regulation in childhood as a predictor of future outcomes: A meta-analytic review. *Psychological Bulletin, 146*(5), 324-354. http://dx.doi.org/10.1037/bul0000227

Slavin, R. (2017, September 28). *Little sleepers: Long-term effects of preschool.* Robert Slavin's Blog. https://robertslavinsblog.wordpress.com/2017/09/28/little-sleepers-long-term-effects-of-preschool/

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35*, 401-426. https://doi.org/10.1016/j.intell.2006.09.004

Von Hippel, P. A., Chor, E., Lurie, L. (2025). *Is Head Start worth saving?* Education Next. https://www.educationnext.org/is-head-start-worth-saving-project-2025-eliminate-program-research

Wolchik, S. A., Tein, J.-Y., Rhodes, C. A., Sandler, I. N., Luecken, L. J., & Porter, M. M. (2025). Cascade effects of a parenting-focused program for divorced families on three health-related outcomes in emerging adulthood. *Development and Psychopathology, 37*(2), 1027–1043. https://doi.org/10.1017/S0954579424000889

Zhang, N., Sandler, I., Tein, J.-Y., & Wolchik, S. (2023). Reducing suicide risk in parentally bereaved youth through promoting effective parenting: Testing a developmental cascade model. *Development and Psychopathology, 35*(1), 433–446. https://doi.org/10.1017/S0954579421001474

References for Papers Included in the Sample

Early Training Project (ETP)

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association, 103*(484), 1481-1495. https://doi.org/10.1198/016214508000000841

Klaus R. A., & Gray S. W. (1986). The Early Training Project for disadvantaged children: A report after five years. *Monographs of the Society in Child Development, 33*(4), 1-66.

Perry Preschool

Elango, S., García, J. L., Heckman, J. J., & Hojman, A. (2015). Early childhood education. *NBER Working Paper* 21766. https://doi.org/10.3386/w21766

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

Weikart D. P., Deloria, D. J., & Lawser, S. A. (1970). *Longitudinal Results of the Ypsilanti Perry Preschool Project. Final Report.* Department of Health, Education, and Welfare.

Abecedarian Preschool Project

Campbell, F. A., & Ramey, C. T. (1990). The relationship between Piagetian cognitive development, mental test performance, and academic achievement in high-risk students with and without early educational experience. *Intelligence, 14*(3), 293-308. https://doi.org/10.1016/0160-2896(90)90020-T

Elango, S., García, J. L., Heckman, J. J., & Hojman, A. (2015). Early childhood education. *NBER Working Paper* 21766. https://doi.org/10.3386/w21766

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

Pages, R., Protzko, J., & Bailey, D. H. (2022). The breadth of impacts from the Abecedarian Project early intervention on cognitive skills. *Journal of Research on Educational Effectiveness, 15* (2), 243-262.

Nurse Home Visitation Program (NHVP)

Eckenrode, J., Campa, M., Luckey, D.W., Henderson, C. R., Cole, R., Kitsman, H., Arson, E., Sidora, K., Powers, P., & Olds, D. L. (2010) Long-term effects of prenatal and infancy nurse home visitation on the life course of youths 19-Year follow-up of a randomized trial. A*rchives of Pediatric Adolescent Medicine, 164*(1), 9–15. https://doi.org/10.1001/archpediatrics.2009.240

Olds, D. L., Henderson, C. R., Chamberlin, R., & Tatelbaum, R. (1986). Preventing child abuse and neglect: A randomized trial of nurse home visitation. *Pediatrics, 78*(1), 65–78. https://doi.org/10.1542/peds.78.1.65

Olds, D. L., Henderson, C. R., & Kitzmna, H. (1994). Does prenatal and infancy nurse home visitation have enduring effects on qualities of parental caregiving and child health at 25 to 50 months of life? *Pediatrics, 93*(1), 89–98. https://doi.org/10.1542/peds.93.1.89

Project Carolina Approach to Responsive Education (CARE) ~ Center-based plus Family Education Group

Campbell, F. A., Wasik, B. H., Pungello, E., Burchinal, M., Barbarin, O., Kainz, K., Sparling, J. J., Ramey, C. T. (2008). Young adult outcomes of the Abecedarian and CARE early childhood educational interventions. *Early Childhood Research Quarterly, 23*(4), 452–466. https://doi.org/10.1016/j.ecresq.2008.03.003

Campbell, F. A., Conti, G., Heckman, J. J., Moon, S. H., & Pinto, R. (2013). The effects of early intervention on human development and social outcomes: Provisional evidence from ABC and CARE. *University of Chicago, Department of Economics*.

Project Carolina Approach to Responsive Education (CARE) ~ Family Education Group

Campbell, F. A., Wasik, B. H., Pungello, E., Burchinal, M., Barbarin, O., Kainz, K., Sparling, J. J., Ramey, C. T. (2008). Young adult outcomes of the Abecedarian and CARE early childhood educational interventions. *Early Childhood Research Quarterly, 23*(4), 452–466. https://doi.org/10.1016/j.ecresq.2008.03.003

Campbell, F. A., Conti, G., Heckman, J. J., Moon, S. H., & Pinto, R. (2013). The effects of early intervention on human development and social outcomes: Provisional evidence from ABC and CARE. *University of Chicago, Department of Economics*.

Longitudinal Comparison Project (Mediated Learning)

Cole, K. N., Dale, P. S., Mills, P. E., Jenkins, J. R. (1993). Interaction between early intervention curricula and student characteristics. *Exceptional Children, 60*(1), 17-28.

Jenkins, J. R., Dale, P. S., Mills, P. E., Cole, K. N., Pious, C., & Ronk, J. (2006). How special education preschool graduates finish: Status at 19 years of age. *American Educational Research Journal, 43*(4), 737-781. https://doi.org/10.3102/00028312043004737

Infant Health and Development Program (IHDP)

Hill, J. L., Brooks-Gunn, J., & Waldfogel, J. (2003). Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology, 39*(4), 730-744. http://dx.doi.org/10.1037/0012-1649.39.4.730

IHDP (1990). Enhancing the outcomes of low-birth-weight, premature infants. *JAMA, 263*(22), 3035-3042.

McCormick, M. C., Brooks-Gunn, J., Buka, S. L., Goldman, J., Yu, J., Salganik, M., Scott, D. T., Bennett, F. C., Kay, L. L., Bernbaum, J. C., Bauer, C. R., Martin, C., Woods, E. R., Martin, A., & Casey, P. H. (2006). Early intervention in low birth weight premature infants: results at 18 years of age for the Infant Health and Development Program. *Pediatrics, 117*(3), 771-780. https://doi.org/10.1542/peds.2005-1316

Project Star

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

Dynarski, S., Hyman, J., & Schanzenbach, D. W. (2013). Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion. *Journal of Policy Analysis and Management*, *32*(4), 692-717. https://doi.org/10.1002/pam.21715

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, *114*(2), 497-532.

Muennig, P., Johnson, G., & Wilde, E. T. (2011). The effect of small class sizes on mortality through age 29 years: evidence from a multicenter randomized controlled trial. *American journal of epidemiology*, *173*(12), 1468-1474. https:/doi.org/10.1093/aje/kwr011

Schanzenbach, D. W. (2006). What have researchers learned from Project STAR?. *Brookings Papers on Education Policy*, (9), 205-228.

Wilde, E. T., Finn, J., Johnson, G., & Muennig, P. (2011). The effect of class size in grades K-3 on adult earnings, employment, and disability status: evidence from a multi-center randomized controlled trial. *Journal of Health Care for the Poor and Underserved*, *22*(4), 1424-1435. https://doi.org/10.1353/hpu.2011.0148

Good Behavior Game 2 (GBG 2)

Dolan, L. J., Kellam, S. G., Brown, C. H., Werthamer-Larsson, L., Rebok, G. W., Mayer, L. S., Laudolff, J., Turkkan, J. S., Ford, C., & Wheeler, L. (1993). The short-term impact of two classroom-based preventive interventions on aggressive and shy behaviors and poor achievement. *Journal of Applied Developmental Psychology*, *14*(3), 317–345. https://doi.org/10.1016/0193-3973(93)90013-L

Poduska, J. M., Kellam, S. G., Wang, W., Brown, C. H., Ialongo, N. S., & Toyinbo, P. (2008). Impact of the Good Behavior Game, a universal classroom-based behavior intervention, on young adult service use for problems with emotions, behavior, or drugs or alcohol. *Drug and Alcohol Dependence*, *95*, S29-S44. https://doi.org/10.1016/j.drugalcdep.2007.10.009

Kellam, S. G., Brown, C. H., Poduska, J. M., Ialongo, N. S., Wang, W., Toyinbo, P., Petras, H., Ford, C., Windham, A., & Wilcox, H. C. (2008). Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. *Drug and Alcohol Dependence*, *95*, S5–S28. https://doi.org/10.1016/j.drugalcdep.2008.01.004

Early Home Visiting Program (EHVP)

Conti, G., Smith, J., Anson, E., Groth, S., Knudtson, M., Salvati, A., & Olds, D. (2024). Early home visits and health outcomes in low-income mothers and offspring 18-year follow-up of a randomized clinical trial. *JAMA Network Open, 7*(1), 1-15. https://doi.org/10.1001/jamanetworkopen.2023.51752

Donelan-McCall, N. S., Knudtson, M. D., & Olds, D. L. (2021). Maternal and child mortality: Analysis of nurse home visiting in 3 RCTs. *American Journal of Preventive Medicine*, *61*(4), 483-491. https://doi.org/10.1016/j.amepre.2021.04.014

Kitzman, H., Olds, D. L., Henderson, C. R., Hanks, C., Cole, R., Tatelbaum, R., ... & Barnard, K. (1997). Effect of prenatal and infancy home visitation by nurses on pregnancy outcomes, childhood injuries, and repeated childbearing: a randomized controlled trial. *JAMA*, *278*(8), 644-652.

Kitzman, H., Olds, D. L., Knudtson, M. D., Cole, R., Anson, E., Smith, J. A., ... & Conti, G. (2019). Prenatal and infancy nurse home visiting and 18-year outcomes of a randomized trial. *Pediatrics*, *144*(6). https://doi.org/10.1542/peds.2018-3876

Olds, D. L., Kitzman, H., Knudtson, M. D., Anson, E., Smith, J. A., & Cole, R. (2014). Effect of home visiting by nurses on maternal and child mortality: results of a 2-decade follow-up of a randomized clinical trial. *JAMA Pediatrics*, *168*(9), 800-806. https://doi.org/doi:10.1001/jamapediatrics.2014.472

Multimodal Treatment Study of ADHD (MTA)

Hechtman, L., Etcovitch, J., Platt, R., Arnold, L. E., Abikoff, H. B., Newcorn, J. H., ... & Wigal, T. (2005). Does multimodal treatment of ADHD decrease other diagnoses?. *Clinical Neuroscience Research*, *5*(5-6), 273-282. https://doi.org/10.1016/j.cnr.2005.09.007

Jensen, P. S., Arnold, L. E., Swanson, J. M., Vitiello, B., Abikoff, H. B., Greenhill, L. L., ... & Hur, K. (2007). 3-year follow-up of the NIMH MTA study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*(8), 989-1002. https://doi.org/10.1097/chi.0b013e3180686d48

Langberg, J. M., Arnold, L. E., Flowers, A. M., Epstein, J. N., Altaye, M., Hinshaw, S. P., ... & Hechtman, L. (2010). Parent-reported homework problems in the MTA study: Evidence for sustained improvement with behavioral treatment. *Journal of Clinical Child & Adolescent Psychology*, *39*(2), 220-233. https://doi.org/10.1080/15374410903532700

MTA Cooperative Group. (1999). A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, *56*(12), 1073-1086.

MTA Cooperative Group. (2004). National Institute of Mental Health Multimodal Treatment Study of ADHD follow-up: 24-month outcomes of treatment strategies for attention-deficit/hyperactivity disorder. *Pediatrics*, *113*(4), 754-761.

Swanson, J. M., Arnold, L. E., Molina, B. S., Sibley, M. H., Hechtman, L. T., Hinshaw, S. P., ... & Stern, K. (2017). Young adult outcomes in the follow-up of the multimodal treatment study of attention-deficit/hyperactivity disorder: Symptom persistence, source discrepancy, and height suppression. *Journal of Child Psychology and Psychiatry*, *58*(6), 663-678. https://doi.org/10.1111/jcpp.12684

Wells, K. C., Chi, T. C., Hinshaw, S. P., Epstein, J. N., Pfiffner, L., Nebel-Schwalm, M., ... & Wigal, T. (2006). Treatment-related changes in objectively measured parenting behaviors in the multimodal treatment study of children with attention-deficit/hyperactivity disorder. *Journal of Consulting and Clinical Psychology*, *74*(4), 649. https://doi.org/10.1037/0022-006X.74.4.649

New Beginnings ~ Mother-only and dual-component groups

Mahrer, N. E., Winslow, E., Wolchik, S. A., Tein, J. Y., & Sandler, I. N. (2014). Effects of a preventive parenting intervention for divorced families on the intergenerational transmission of parenting attitudes in young adult offspring. *Child Development*, *85*(5), 2091-2105. https://doi.org/10.1111/cdev.l2258

Rhodes, C. A. A. (2019). *Intervention Effects on Coping and Coping Efficacy: A Fifteen-Year Follow-Up of the New Beginnings Program* (Master's thesis, Arizona State University).

Sigal, A. B., Wolchik, S. A., Tein, J. Y., & Sandler, I. N. (2012). Enhancing youth outcomes following parental divorce: A longitudinal study of the effects of the New Beginnings Program on educational and occupational goals. *Journal of Clinical Child & Adolescent Psychology*, *41*(2), 150-165. https://doi.org/10.1080/15374416.2012.651992

Vélez, C. E., Wolchik, S. A., Tein, J. Y., & Sandler, I. (2011). Protecting children from the consequences of divorce: A longitudinal study of the effects of parenting on children's

coping processes. *Child Development*, *82*(1), 244-257. https://doi.org/10.1111/j.!467-8624.2010.01553.x
Wolchik, S. A., Sandler, I. N., Tein, J. Y., Mahrer, N. E., Millsap, R. E., Winslow, E., ... & Reed, A. (2013). Fifteen-year follow-up of a randomized trial of a preventive intervention for divorced families: effects on mental health and substance use outcomes in young adulthood. *Journal of Consulting and Clinical Psychology*, *81*(4), 660. https://doi.org/10.1037/a0033235
Wolchik, S. A., Tein, J. Y., Winslow, E., Minney, J., Sandler, I. N., & Masten, A. S. (2021). Developmental cascade effects of a parenting-focused program for divorced families on competence in emerging adulthood. *Development and Psychopathology*, *33*(1), 201-215. https://doi.org/10.1017/S095457941900169X

Good Behavior Game (GBG)~ Classroom
Bradshaw, C. P., Zmuda, J. H., Kellam, S. G., & Ialongo, N. S. (2009). Longitudinal impact of two universal preventive interventions in first grade on educational outcomes in high school. *Journal of Educational Psychology*, *101*(4), 926. https://doi.org/10.1037/a0016586
Ialongo, N. S., Werthamer, L., Kellam, S. G., Brown, C. H., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. *American Journal of Community Psychology*, *27*(5), 599-641.
Musci, R. J., Fairman, B., Masyn, K. E., Uhl, G., Maher, B., Sisto, D. Y., ... & Ialongo, N. S. (2018). Polygenic score× intervention moderation: an application of discrete-time survival analysis to model the timing of first marijuana use among urban youth. *Prevention Science*, *19*(1), 6-14. https://doi.org/10.1007/s11121-016-0729-1
Saunders, J. M. (2007). *An empirical test of Terrie Moffitt's developmental taxonomy of delinquency*. City University of New York.
Wang, Y., Storr, C. L., Green, K. M., Zhu, S., Stuart, E. A., Lynne-Landsman, S. D., ... & Ialongo, N. S. (2012). The effect of two elementary school-based prevention interventions on being offered tobacco and the transition to smoking. *Drug and Alcohol Dependence*, *120*(1-3), 202-208. https://doi.org/10.1016/j.drugalcdep.2011.07.022

Good Behavior Game (GBG)~ Family
Bradshaw, C. P., Zmuda, J. H., Kellam, S. G., & Ialongo, N. S. (2009). Longitudinal impact of two universal preventive interventions in first grade on educational outcomes in high school. *Journal of Educational Psychology*, *101*(4), 926. https://doi.org/10.1037/a0016586
Ialongo, N. S., Werthamer, L., Kellam, S. G., Brown, C. H., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. *American Journal of Community Psychology*, *27*(5), 599-641.
Musci, R. J., Fairman, B., Masyn, K. E., Uhl, G., Maher, B., Sisto, D. Y., ... & Ialongo, N. S. (2018). Polygenic score× intervention moderation: an application of discrete-time survival analysis to model the timing of first marijuana use among urban youth. *Prevention Science*, *19*(1), 6-14. https://doi.org/10.1007/s11121-016-0729-1

Saunders, J. M. (2007). *An empirical test of Terrie Moffitt's developmental taxonomy of delinquency*. City University of New York.

Wang, Y., Storr, C. L., Green, K. M., Zhu, S., Stuart, E. A., Lynne-Landsman, S. D., ... & Ialongo, N. S. (2012). The effect of two elementary school-based prevention interventions on being offered tobacco and the transition to smoking. *Drug and Alcohol Dependence*, *120*(1-3), 202-208. https://doi.org/10.1016/j.drugalcdep.2011.07.022

Project Family ~ Preparing for the Drug-Free Years Group

Mason, W. A., Kosterman, R., Hawkins, J. D., Haggerty, K. P., & Spoth, R. L. (2003). Reducing adolescents' growth in substance use and delinquency: Randomized trial effects of a parent-training prevention intervention. *Prevention Science*, *4*, 203-212.

Mason, W. A., Kosterman, R., Hawkins, J. D., Haggerty, K. P., Spoth, R. L., & Redmond, C. (2007). Influence of a family-focused substance use preventive intervention on growth in adolescent depressive symptoms. *Journal of Research on Adolescence*, *17*(3), 541-564.

Mason, W. A., Kosterman, R., Haggerty, K. P., Hawkins, J. D., Redmond, C., Spoth, R. L., & Shin, C. (2009). Gender moderation and social developmental mediation of the effect of a family-focused substance use preventive intervention on young adult alcohol abuse. *Addictive Behaviors*, *34*(6-7), 599-605. https://doi.org/10.1016/j.addbeh.2009.03.032

Park, J., Kosterman, R., Hawkins, J. D., Haggerty, K. P., Duncan, T. E., Duncan, S. C., & Spoth, R. (2000). Effects of the "Preparing for the Drug Free Years" curriculum on growth in alcohol use and risk for alcohol use in early adolescence. *Prevention Science*, *1*, 125-138.

Spoth, R. L., Clair, S., Shin, C., & Redmond, C. (2006). Long-term effects of universal preventive interventions on methamphetamine use among adolescents. *Archives of Pediatrics & Adolescent Medicine*, *160*(9), 876-882.

Spoth, R., Trudeau, L., Shin, C., & Redmond, C. (2008a). Long-term effects of universal preventive interventions on prescription drug misuse. *Addiction*, *103*(7), 1160-1168. https://doi.org/10.1111/j.1360-0443.2008.02160.x

Spoth, R., Trudeau, L., Guyll, M., Shin, C., & Redmond, C. (2009a). Universal intervention effects on substance use among young adults mediated by delayed adolescent substance initiation. *Journal of Consulting and Clinical Psychology*, *77*(4), 620. https://doi.org/10.1037/a0016029

Project Family ~ Strengthening Families Group

Spoth, R. L., Clair, S., Shin, C., & Redmond, C. (2006). Long-term effects of universal preventive interventions on methamphetamine use among adolescents. *Archives of Pediatrics & Adolescent Medicine*, *160*(9), 876-882.

Spoth, R., Guyll, M., & Shin, C. (2009). Universal intervention as a protective shield against exposure to substance use: Long-term outcomes and public health significance. *American Journal of Public Health*, *99*(11), 2026-2033. https://doi.org/10.2105/AJPH.2007.133298

Spoth, R., Guyll, M., Trudeau, L., & Goldberg-Lillehoj, C. (2002). Two studies of proximal outcomes and implementation quality of universal preventive interventions in a community–university collaboration context. *Journal of Community Psychology*, *30*(5), 499-518. https://doi.org/10.1002/jcop.10021

Spoth, R., Randall, G. K., & Shin, C. (2008). Increasing school success through partnership-based family competency training: Experimental study of long-term outcomes. *School Psychology Quarterly*, *23*(1), 70. https://doi.org/10.1037/1045-3830.23.1.70

Spoth, R., Redmond, C., & Lepper, H. (1999). Alcohol initiation outcomes of universal family-focused preventive interventions: one-and two-year follow-ups of a controlled study. *Journal of Studies on Alcohol*, (13), 103-111.

Spoth, R., Trudeau, L., Shin, C., & Redmond, C. (2008a). Long-term effects of universal preventive interventions on prescription drug misuse. *Addiction*, *103*(7), 1160-1168. https://doi.org/10.1111/j.1360-0443.2008.02160.x

Spoth, R., Trudeau, L., Guyll, M., Shin, C., & Redmond, C. (2009a). Universal intervention effects on substance use among young adults mediated by delayed adolescent substance initiation. *Journal of Consulting and Clinical Psychology*, *77*(4), 620. https://doi.org/10.1037/a0016029

Spoth, R., Trudeau, L., Shin, C., Ralston, E., Redmond, C., Greenberg, M., & Feinberg, M. (2013). Longitudinal effects of universal preventive intervention on prescription drug misuse: three randomized controlled trials with late adolescents and young adults. *American Journal of Public Health*, *103*(4), 665-672. https://doi.org/10.2105/AJPH. 2012.301209

Trudeau, L., Spoth, R., Randall, G. K., Mason, W. A., & Shin, C. (2012). Internalizing symptoms: Effects of a preventive intervention on developmental pathways from early adolescence to young adulthood. *Journal of Youth and Adolescence*, *41*, 788-801. https://doi.org/10.1007/s10964-011-9735-6

Project Alliance #1

Gardner, T. W., Dishion, T. J., & Connell, A. M. (2008). Adolescent self-regulation as resilience: Resistance to antisocial behavior within the deviant peer context. *Journal of Abnormal Child Psychology*, *36*, 273-284. https://doi.org/10.1007/s10802-007-9176-6

Panza, K. E. (2019). High-Risk Sexual Behavior and Substance Use During Young Adulthood: Gender-Specific Developmental Trajectories and the Influence of Early Trauma, and Adolescent Peer and Family Processes. *Arizona State University*.

Van Ryzin, M. J., & Nowicka, P. (2013). Direct and indirect effects of a family-based intervention in early adolescence on parent− youth relationship quality, late adolescent health, and early adult obesity. *Journal of Family Psychology*, *27*(1), 106. https://doi.org/10.1037/a0031428

Van Ryzin, M. J., & Dishion, T. J. (2012). The impact of a family-centered intervention on the ecology of adolescent antisocial behavior: Modeling developmental sequelae and trajectories during adolescence. *Development and Psychopathology*, *24*(3), 1139-1155. https://doi.org/10.1017/S0954579412000582

Zhang, J., Hanson, A. N., Piehler, T. F., & Ha, T. (2024). Coercive parent-adolescent interactions predict substance use and antisocial behaviors through early adulthood: A dynamic systems perspective. *Research on Child and Adolescent Psychopathology*, *52*(1), 141-154. https://doi.org/10.1007/s10802-023-01102-8

Reading Remediation

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

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

Capable Families and Youth Strengthening Families Program

Spoth, R. L., Clair, S., Shin, C., & Redmond, C. (2006). Long-term effects of universal preventive interventions on methamphetamine use among adolescents. *Archives of Pediatrics & Adolescent Medicine*, *160*(9), 876-882.

Spoth, R., Guyll, M., Trudeau, L., & Goldberg-Lillehoj, C. (2002). Two studies of proximal outcomes and implementation quality of universal preventive interventions in a community–university collaboration context. *Journal of Community Psychology*, *30*(5), 499-518. https://doi.org/10.1002/jcop.10021

Spoth, R., Randall, G. K., Shin, C., & Redmond, C. (2005). Randomized study of combined universal family and school preventive interventions: patterns of long-term effects on initiation, regular use, and weekly drunkenness. *Psychology of Addictive behaviors*, *19*(4), 372. https://doi.org/10.1037/0893-164x.19.4.372

Spoth, R., Trudeau, L., Shin, C., Ralston, E., Redmond, C., Greenberg, M., & Feinberg, M. (2013). Longitudinal effects of universal preventive intervention on prescription drug misuse: three randomized controlled trials with late adolescents and young adults. *American Journal of Public Health*, *103*(4), 665-672. https://doi.org/10.2105/AJPH.2012.301209

Spoth, R., Trudeau, L., Shin, C., & Redmond, C. (2008). Long-term effects of universal preventive interventions on prescription drug misuse. *Addiction*, *103*(7), 1160-1168. https://doi.org/10.1111/j.1360-0443.2008.02160.x

Trudeau, L., Spoth, R., Lillehoj, C., Redmond, C., & Wickrama, K. A. (2003). Effects of a preventive intervention on adolescent substance use initiation, expectancies, and refusal intentions. *Prevention Science*, *4*, 109-122.

Trudeau, L., Spoth, R., Mason, W. A., Randall, G. K., Redmond, C., & Schainker, L. (2016). Effects of adolescent universal substance misuse preventive interventions on young adult depression symptoms: Mediational modeling. *Journal of Abnormal Child Psychology*, *44*, 257-268.https://doi.org/10.1007/s10802-015-9995-9

PROSPER Family- and School-focused Intervention

Redmond, C., Spoth, R. L., Shin, C., Schainker, L. M., Greenberg, M. T., & Feinberg, M. (2009). Long-term protective factor outcomes of evidence-based interventions implemented by community teams through a community–university partnership. *The Journal of Primary Prevention*, *30*, 513-530. https://doi.org/10.1007/s10935-009-0189-5

Spoth, R., Redmond, C., Clair, S., Shin, C., Greenberg, M., & Feinberg, M. (2011). Preventing substance misuse through community–university partnerships: Randomized controlled trial outcomes 4½ years past baseline. *American Journal of Preventive Medicine*, *40*(4), 440-447. https://doi.org/10.1016/j.amepre.2010.12.012

Spoth, R., Trudeau, L., Shin, C., Ralston, E., Redmond, C., Greenberg, M., & Feinberg, M. (2013). Longitudinal effects of universal preventive intervention on prescription drug misuse: three randomized controlled trials with late adolescents and young adults. *American Journal of Public Health*, *103*(4), 665-672. https://doi.org/10.2105/AJPH.2012.301209

Spoth, R., Redmond, C., Shin, C., Greenberg, M. T., Feinberg, M. E., & Trudeau, L. (2017). PROSPER delivery of universal preventive interventions with young adolescents: long-term effects on emerging adult substance misuse and associated risk behaviors. *Psychological Medicine*, *47*(13), 2246-2259. https://doi.org/10.1017/S0033291717000691

Spoth, R., Redmond, C., Shin, C., Trudeau, L., Greenberg, M. T., Feinberg, M. E., & Welsh, J. (2022). Applying the PROSPER prevention delivery system with middle schools: Emerging adulthood effects on substance misuse and conduct problem behaviors through 14 years past baseline. *Child Development*, *93*(4), 925-940. https://doi.org/10.1111/cdev.13746

Chicago School Readiness Project (CSRP)

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Metzger, M. W., Solomon, B. (2009). Targeting children's behavior problems in preschool classrooms: A cluster-randomized controlled trial. *Journal of Consulting and Clinial Psychology, 77*(2), 302-316. https://doi.org/10.1037/a0015302

Raver, C. C., Li-Grining, C., Bub, K., Jones, S. M., Zhai, F., & Presller, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism. *Child Development, 82*(1), 362-378. https://doi.org/10.1111/j.1467-8624.2010.01561.x

Watts, T. W., Li, C., Pan, X. S., Gandhi, J., McCoy, D. C., & Raver, C. C. (2023). Impacts of the Chicago School Readiness Project on measures of achievement, cognitive functioning, and behavioral regulation in late adolescence. *Developmental Psychology, 59* (12), 2204-2222. https://doi.org/10.1037/dev0001561

Charter Middle Schools

Clark, M. A., Gleason, P. M., Clark Tuttle, C., & Silverberg, M. K. (2015). Do charter schools improve student achievement? *Educational Evaluation and Policy Analysis, 37*(4), 419-436. https://doi.org/10.3102/0162373714558292

NCEEA. (2019). *Do Charter Middle Schools Improve Students' College Outcomes?* IES National Center for Educational Evaluation and Regional Assistance.

Higher Achievement

Herrera, C., Grossman, J. B., Linen, L. L. (2013). *Staying on track: Testing Higher Achievement's long-term impact on academic outcomes and high school choice.* MDRC.

Garcia, I., Grossman, J. B., Herrera, C. Linden, L. L. (2020). *The Impact of an Intensive Year-Round Middle School Program on College Attendance.* MDRC.

Knowledge is Power Program (KIPP)

Coen, T., Nichols-Barrer, I., Gleason, P. (2019). *Long-term Impacts of KIPP Middle Schools on College Enrollment and Early College Persistence.* Mathematica.

Demers, A., Nichols-Barrer, I., Steele, E., Bartlett, M. & Gleason, P. (2023). *Long-term Impacts* *of KIPP Middle and High Schools on College Enrollment, Persistence, and Attainment.* Mathematica.

Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I. & Resch, A. (2013). *KIPP Middle Schools: Impacts on Achievement and Other Outcomes. Final Report.* Mathematica.

Project Alliance #2

Danzo, S., Connell, A. M., & Stormshak, E. (2021). Pathways between alcohol use and internalizing symptoms across emerging adulthood: Examination of gender differences in interpersonal and intrapersonal processes. *Emerging Adulthood*, *9*(4), 347-359. https://doi.org/10.1177/2167696820936066

Fosco, G. M., Frank, J. L., Stormshak, E. A., & Dishion, T. J. (2013). Opening the "Black Box": Family Check-Up intervention effects on self-regulation that prevents growth in problem behavior and substance use. *Journal of School Psychology*, *51*(4), 455-468. http://dx.doi.org/10.1016/j.jsp.2013.02.001

Fosco, G. M., Van Ryzin, M. J., Connell, A. M., & Stormshak, E. A. (2016). Preventing adolescent depression with the family check-up: Examining family conflict as a mechanism of change. *Journal of Family Psychology*, *30*(1), 82. http://dx.doi.org/10.1037/fam0000147

Stormshak, E., DeGarmo, D., Chronister, K., & Caruthers, A. (2018). The impact of family-centered prevention on self-regulation and subsequent long-term risk in emerging adults. *Prevention Science*, *19*, 549-558. https://doi.org/10.1007/s11121-017-0852-7

Staying Connected ~ Parent and Adolescent Administered

Haggerty, K. P., Skinner, M. L., MacKenzie, E. P., & Catalano, R. F. (2007). A randomized trial of parents who care: Effects on key outcomes at 24-month follow-up. *Prevention Science, 8*, 249-260. https://doi.org/10.1007/s11121-007-0077-2

Haggerty, K. P., Klima, T., Skinner, M. L., Catalano, R. F., & Barkan, S. (2015). Staying Connected with your Teen and the promise of self-directed prevention programs. In *Family-Based Prevention Programs for Children and Adolescents* (pp. 221-240). Psychology Press.

Staying Connected ~ Self-administered with Telephone Support

Haggerty, K. P., Skinner, M. L., MacKenzie, E. P., & Catalano, R. F. (2007). A randomized trial of parents who care: Effects on key outcomes at 24-month follow-up. *Prevention Science, 8*, 249-260. https://doi.org/10.1007/s11121-007-0077-2

Haggerty, K. P., Klima, T., Skinner, M. L., Catalano, R. F., & Barkan, S. (2015). Staying Connected with your Teen and the promise of self-directed prevention programs. In *Family-Based Prevention Programs for Children and Adolescents* (pp. 221-240). Psychology Press.

Strong African American Families (SAAF)

Brody, G. H., Kogan, S. M., Chen, Y., & McBride Murry, V. (2008). Long-term effects of the Strong African American Families Program on Youths' Conduct Problems. *Journal of Adolescent Health, 43*, 474-481.

Miller G. E., Brody, G. H., Yu, T., & Chen, E. (2014). A family-oriented psychosocial intervention reduces inflammation in low-SES African American youth. *PNAS, 111*(31), 11287-11292. https://doi.org/10.1073/pnas.1406578111