EdWorkingPaper No. 25-1367

Childhood Interventions and Life Course Development

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 paradox has perplexed researchers studying childhood interventions: although program impacts on children's skills often fade, some interventions have nonetheless produced long-run impacts on adult outcomes. Existing developmental theory does not provide a straightforward explanation. The fadeout-emergence paradox spotlights our limited understanding of how early skill gains shape long-run developmental trajectories. The current meta-analysis examined the longitudinal impacts of educational and developmental randomized controlled trials that measured initial impacts on child skills and long-run impacts on adult outcomes. We identified a diverse sample of 29 interventions from 25 studies from which we analyzed 179 posttest effects and 497 adult effects. We found that initial impacts on child skills faded considerably in the two years after interventions ended. Nonetheless, some programs generated long-run effects on adult outcomes that were of a similar magnitude to the faded effects (~.04 SD, 95% PI [.00, .09]). We observed some, but not very strong, correspondence between initial intervention impacts on child skills and long-run impacts on adult outcomes. Our findings illustrate substantial gaps in our knowledge of the mechanisms linking childhood skills to adult functioning. Skill-building theories that hinge on the persistence of targeted skills require revision. We articulate recommendations to advance intervention science.

VERSION: December 2025

Suggested citation: Hart, Emma R., Drew H. Bailey, Casey Moran, Susan Kruglinski, and Tyler W. Watts. (2025). Childhood Interventions and Life Course Development. (EdWorkingPaper: 25-1367). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/dtk5-ds56

Draft version 1, 11/19/25.
This paper has not been peer reviewed.

Childhood Interventions and Life Course Development

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.

Abstract

A paradox has perplexed researchers studying childhood interventions: although program impacts on children's skills often fade, some interventions have nonetheless produced long-run impacts on adult outcomes. Existing developmental theory does not provide a straightforward explanation. The fadeout-emergence paradox spotlights our limited understanding of how early skill gains shape long-run developmental trajectories. The current meta-analysis examined the longitudinal impacts of educational and developmental randomized controlled trials that measured initial impacts on child skills and long-run impacts on adult outcomes. We identified a diverse sample of 29 interventions from 25 studies from which we analyzed 179 posttest effects and 497 adult effects. We found that initial impacts on child skills faded considerably in the two years after interventions ended. Nonetheless, some programs generated long-run effects on adult outcomes that were of a similar magnitude to the faded effects (~.04 *SD*, 95% PI [.00, .09]). We observed some, but not very strong, correspondence between initial intervention impacts on child skills and long-run impacts on adult outcomes. Our findings illustrate substantial gaps in our knowledge of the mechanisms linking childhood skills to adult functioning. Skill-building theories that hinge on the persistence of targeted skills require revision. We articulate recommendations to advance intervention science.

Keywords: educational and developmental randomized controlled trials, long-run effects, fadeout, meta-analysis, developmental theory

Childhood Interventions and Life Course Development

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

In 2023, Gray-Lobe and colleagues (2023) published an evaluation of the longitudinal effects of Boston’s preschool program. The study joined a small yet influential literature on when and how educational interventions yield long-run benefits. Gray-Lobe’s findings were methodologically rigorous and theoretically perplexing: children randomized to receive a year of preschool demonstrated higher educational attainment in adulthood than children who did not receive preschool, even in the absence of group differences on elementary school test scores and special-education placement. In providing a rare glimpse into the “long arm” of childhood, longitudinal intervention evaluations with counterintuitive results like Gray-Lobe’s have piqued the interest of researchers, policymakers, and the public alike.

The study suggested the possibility of *emergence*: that educational experiences could generate effects on life course outcomes despite null effects on hypothesized child skill mediators. How could this be possible?

Similarly mysterious patterns have been observed across the psychological literature and are generally described as “sleeper effects.” This term has taken on many meanings in psychology and even popular culture. It is often used colloquially for any effect that appears unexpectedly at a later point, such as when an initially low-profile movie gradually emerges as a fan-favorite (e.g., "That film was a sleeper."). The concept was introduced into psychology about 75 years ago when Hovland and colleagues (1949) found that although individuals initially discounted information provided by untrustworthy sources, some eventually came to accept it, much to the researchers' surprise. Social psychologists proceeded to investigate whether the apparent sleeper effect in beliefs was real and reflective of a "generalized phenomenon" (Capon & Hulbert, 1973).

The field of human development has produced several studies with perplexing patterns of "sleeper" effects in which early experiences are particularly predictive of outcomes measured much later in life (Clarke & Clarke, 1981; Sarigiani & Spierling, 2011; Thompson et al., 2024). Theoretical explanations abound. Kagan (1971) articulated "heterotypic continuity" as a regularity of development wherein core underlying processes result in different behavioral expressions across stages of development. Perhaps ecological contexts like school lead children to inhibit traits that then re-express later in life (Kagan & Moss, 1962), or the competencies developed in a supportive early context only appear during later-life transitions (Bronfenbrenner, 1979). Even Freud (1922) argued that childhood experiences can lead to neuroses that will only become apparent in adulthood.

Many of these theories have been met with strong skepticism (Capon & Hulbert, 1973; Clarke & Clarke, 1981; Cook et al., 1979). Sleeper effects may appear because we lack good measures of mediating mechanisms that lead interventions to impact adult outcomes. Replication is also a concern; famous examples of sleeper effects might be unlikely to replicate due to the use of outdated methodologies, including small sample sizes.

Yet, Gray-Lobe and colleagues' (2023) preschool findings are not without precedent in the modern educational intervention literature. A robust examination of random assignment to classrooms with varying teacher quality found long-run impacts on adult outcomes despite the fadeout of initial impacts on achievement (Chetty et al., 2011). Deming's (2009) analysis of

Head Start impacts also found fadeout on cognitive skills throughout the elementary and middle school years, followed by meaningful effects on adult outcomes. Other studies have suggested that end-of-treatment differences between children who received an intervention versus those who did not may provide little indication of the differences that will be present in the long-run. For example, Hitt et al. (2018) observed that although school choice programs could have long-term effects, the short-term effectiveness of such programs appeared to be wholly unrelated to those long-term effects.

While grappling with the possibility of "sleeper effects" from educational programs, the field has also accumulated evidence that educational and developmental interventions often produce effects that fade out in the medium-term (Bailey et al., 2017; Bailey et al., 2020; Bus & van IJzendoorn, 1999; Hart et al., 2024; Hattie et al., 1996; Protzko, 2015; Rosengarten et al., 2024). Because most studies do not conduct adult follow-up assessments, in most cases where medium-term fadeout has been reported, the accompanying long-run adult effects remains unknown (Watts et al., 2019).

The possibility of fadeout followed by emergent adult impacts is challenging to square with existing developmental theory (e.g., Masten & Cicchetti, 2010). Why would later impacts emerge if effects on targeted psychological constructs were short-lived? This puzzle raises concerns about the science of interventions, namely whether research can inform educational investments that will reliably yield long-run benefits. Indeed, the field has long struggled with the concern that the effects of educational programs are theoretically underdetermined (Cronbach, 1969). If we do not understand the mediational processes that link our interventions to longer-term effects, then our developmental theories may miss the mark, and new programs may be less likely to reliably benefit children.

To bring new clarity to this complicated, yet important, area of research, we gathered and meta-analyzed a sample of randomized controlled trials (RCTs) that have assessed both short-term effects (i.e., effects measured at the immediate end of the intervention) on child skills and long-run effects on adult outcomes. We additionally coded and analyzed follow-up effects on child skills measured six months to two years after interventions ended, which we refer to as "medium-term" effects. We included programs that targeted child skills prior to high school entry and, thus, arrived at a diverse sample of interventions spanning nurse-home visiting, early childhood education, reading remediation, classroom behavioral management, and adolescent substance prevention. Henceforth, we refer to these programs as "educational" as they attempted to improve children's skills through direct and indirect educational experiences.

Drawing on this dataset, we first set out to clarify the basic descriptive trajectory of longitudinal intervention impacts. Our main objective was to identify whether the fadeout-emergence paradox is broadly observed. We then tested whether short-term impacts on targeted child skills forecasted long-term impacts on adult outcomes. In other words, did programs that generated larger initial effects on child skills also produce larger impacts on adult outcomes? By assessing the relation between short-term effects and adult impacts across an unusually diverse range of educational interventions, it may be possible to better understand the plausibility of the theoretical assumptions underlying current intervention models.

Below, we provide an in-depth review of the relevant literature, beginning with skill-based theories that make predictions about the role of building specific psychological capacities during childhood to support positive adult functioning. We then discuss how these theories have been complicated by recent evidence of intervention fadeout, and newer theoretical explanations that have been offered to square skill-building processes with the reality of fading effects.

Finally, we discuss issues with the current literature that make it difficult to draw clear conclusions about the longer-term effects of educational interventions and motivated our meta-analytic approach for this study.

The Case for Foundational Childhood Skills

Skills Beget Skills

Why would we expect that educational interventions targeting children's skills would affect long-term functioning in the first place? The idea that prior skill acquisition shapes subsequent skill advancements is a central tenet of developmental theory. In the realm of cognitive development, this idea is particularly intuitive. For example, basic decoding skills are thought to enable children to make better use of subsequent reading opportunities in ways that support the development of more advanced literacy skills (Share, 1995). The same logic applies to the development of social-emotional competencies. Schneider-Rosen and Cicchhetti (1984) argued that each stage of development contains distinct social demands that require the acquisition of specific skills and competencies, and a child's success at meeting these demands in their context in part depends on their success or failure at acquiring the requisite skills and competencies at earlier stages. For example, conduct issues at one point in development may increase the likelihood of worsening behavioral issues at a later point in development (Dodge et al., 1986). Prior skills may not only shape subsequent within-domain functioning but also shape functioning in related domains, a process described as *transfer*. Indeed, earlier reading abilities are thought to be necessary, but not sufficient, for the development of more advanced literacy skills. Behavioral regulation skills may additionally impact a child's ability to make progress in reading (Blair & Raver, 2015). Across longer timescales, it is expected that earlier skills can alter developmental trajectories commensurate with variable adult outcomes.

Two particularly influential expressions of these ideas come from Cunha and Heckman's *Technology of Skill Formation* (2007) and Masten and Cicchetti's *Developmental Cascades* (2010). Both theories are consistent with the expectation that differences in early skills will persist, and ultimately come to 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 wherein earlier-developed skills are self-reinforced and sustained within the same domain and across domains. They also argued that dynamic complementarities are a critical mechanism driving skill persistence, whereby stronger skills enable children to make better use of their subsequent learning environments. Indeed, the role of subsequent environmental inputs are essential to theories of skill development; earlier skills are expected to increase the likelihood of later skill advancement in the context of children's post-intervention lives (i.e., contexts, relationships, opportunities).

Developmental psychologists Masten and Cicchetti similarly argued that development occurs through cascades. They articulated that functioning in one domain shapes future functioning in the same and related domains through chain-like reactions. As such, stronger skills at one point in development interact with subsequent environments and experiences, ultimately producing increasing benefits over time through positive feedback loops. Negative feedback loops are also possible, wherein earlier struggles give rise to further issues. Both theories argue that through within-domain and cross-domain interactions, earlier capacities may ultimately come to shape functioning on adult outcomes far down the line.

Most of the empirical support for the causal effects of earlier skills on later outcomes comes from correlational studies. A large body of work has shown strong stability in individuals' skills across time (e.g., Breit et al., 2024; Duncan et al., 2007; Reardon, 2011). Other highly influential correlational work has established links between child skills and adult competencies (Burchinal & Vandell, 2025; Davis-Kean et al., 2022; Koepp et al., 2023; Moffitt et al., 2011; Ritchie & Bates, 2013).

Translation to Interventions

Together with skills-beget-skills theory, correlational results have often motivated the targets, timing, and expectations of interventions. Interventions commonly target skills that are predicted by contextual factors (suggestive of *malleability*) and that are predictive of important long-run outcomes (suggestive of *fundamentality*; see Bailey et al., 2017). The moment in development when a skill becomes predictive of long-run outcomes is often evoked in motivating the timing of programs, as predictiveness is commonly interpreted to signal the influence of environmental inputs that can be intervened on (Bloom, 1964; but also see Zigler & Berman, 1983).

Consider, as an example, the Good Behavior Game. The Good Behavior Game is a classroom behavioral management intervention for first graders. The base program aimed to improve the antecedents of adult substance use and antisocial behaviors—specifically aggressive and shy behaviors—which were correlated with adult psychiatric issues and antisocial behaviors in as early as first grade (Dolan et al., 1993; Ialongo et al., 1999). The program was designed to improve social behaviors through targeting teachers' ability to effectively manage aggression and shyness in the classroom. By targeting teachers' management skills, the program strove to prevent teachers' engagement in the kinds of "coercive" caretaking behaviors that researchers worried caused and exacerbated the emergence of children's aggression at home (Poduska et al., 2008). The investigators anticipated that by changing early behavioral issues around the time that they become predictive of later outcomes, it might be possible to improve childrens' trajectories.

When synthesizing across complex individual intervention theories of change, such as that for the Good Behavior Game, several patterns of longitudinal intervention impact trajectories reoccur which we have attempted to depict in Figure 1. Representing hypothetical trajectories of impacts is very challenging and, thus, our figure is an oversimlication. Indeed, researchers' theories regarding how intervention-generated boosts to child skills might persist (i.e., auto-regressive paths, or self-productivity of children's skills) are generally under-articulated. Beyond the dynamics of child skill persistence, researchers rarely document their expectations regarding how, and to what extent, impacts on child skills will transfer to impacts on adult outcomes (i.e., cross-lagged paths; in the case that later outcomes are skills, sometimes these are called "cross-productivities;" in the case they are economic outcomes, sometimes they are called "returns"). Nonetheless, we believe that the figure is a helpful foundation for considering prevalent theories of skill-building that have informed intervention development.

The most optimistic patterns in Figure 1 have frequently motivated long-run intervention evaluations, perhaps in part because such optimism is required to secure the many resources necessary to conduct long-run intervention evaluations. The patterns depict that intervention-driven boosts to child skills will persist (Figure 1, red "Full Persistence" line) or grow (Figure 1, light blue "Growing Effects" line) and ultimately transfer to domains of adult functioning. Such patterns may be possible under maximized dynamic complementarities and cross-productivities. Correlational work showing less-than-perfect skill stability on the same skills over time may support the expectation of considerable, but not perfect or growing, effects (Figure 1, light green "Considerable Persistence"). Even if auto-regressive and cross-lagged associations among skills are lower than 1:1, the magnitude of such associations are often large enough to support the expectation that earlier boosts will lead to meaningful effects on adult functioning, especially given the benefits of dynamic complementarities and cross-productivities

The Fadeout Problem

Research on the effects of RCT-evaluated interventions that target children's skills often yields results that are at odds with the expectations of developmental theory and related correlational findings. Growing evidence suggests that while children who received an educational intervention tend to have stronger skills than children who did not receive the intervention at program end, benefits tend to fade in the months and years that follow (Bailey et al., 2017, 2020; Bus & van Ikzendoorn, 1999; Hart et al., 2024; Hattie et al., 1996; Protzko, 2015; Rosengarten et al., 2024), and this fadeout often occurs at rates that are much faster than what would be predicted by correlational research (Bailey et al., 2018; Wan et al., 2021). *Fadeout* has raised practical concerns about our ability to implement interventions that generate persistent effects (Bailey et al., 2024) and calls into question assumptions underlying the translation of correlational estimates into long-run intervention-effect forecasts (Bailey et al., 2018; Brick & Bailey, 2020). Indeed, theoretical underdetermination of correlations (i.e., that a correlation could reflect the causal influence of X on Y, or any number of completely different theoretical explanations due to confounding) makes it challenging to accurately attribute the cause of correlations (Duncan et al., 2004; Rohrer & Lucas, 2020).

Skill building theory and findings from the correlational literature would lead us to expect that if there are no differences between children who did, and did not, receive the intervention in the years after the program occurred, then we shouldn't find differences in adulthood (aligned with dark blue "Full Fadeout" in Figure 1). However, this is not always the case. As we described above, a handful of interventions have generated long-run adult impacts despite non-existent initial effects or medium-term fadeout (dark green “Sleeper Effects” or orange “Fadeout-Emergence” in Figure 1; Chetty et al., 2011; Gray-Lobe et al., 2023; Hitt et al., 2018; also see Deming, 2009, though not an RCT). Whether these patterns are reflective of a larger trend across programs is unclear.

Theories to Reconcile Skill Building and Fadeout

The fadeout-emergence pattern raises questions about how educational interventions generate long-run effects if not through persistent impacts on targeted child skills. In discussions about what could be going on, researchers often argue that programs must have generated persistent impacts on unmeasured child skills. Drawing on developmental theory and correlational work (Duckworth et al., 2018; Moffitt et al., 2011), some have claimed that educational interventions are likely to generate adult impacts due to overlooked effects on social-emotional skills (often referred to as “non-cognitive” skills in the economics literature (Chetty et al., 2011; Heckman et al., 2013; Heckman & Kautz, 2012). Until recently, most of the studies showing fadeout have primarily focused on cognitive skills, leaving questions about the dynamics of social-emotional skill persistence in the long term (Abenavoli, 2019). Proponents of social-emotional persistence point to several prominent examples of interventions that generated longer-term benefits, with follow-up analyses providing some indication that longer-term effects may have been partially mediated by social-emotional skill-building (Elango et al., 2016; Schweinhart & Weikart, 1981).

A study using the Meta-Analysis of Educational RCTs with Follow-Up (MERF) recently tested the theory that impacts on social-emotional skills persist more than impacts on cognitive skills and found that they didn’t. Drawing from a sample of 86 educational RCTs that measured intervention effects on child skills at posttest and follow-up, Hart et al. (2024) found that posttest impacts on social-emotional skills faded at a similar rate (~50%) as impacts on cognitive skills in the year after programs ended. At least in the short term, the meta-analytic findings highlighted that intervention impacts on social-emotional skills may be unlikely to drive the long-run effects of interventions (also see Rosengarten et al., 2024).

The findings from MERF suggested that if an intervention generates a long-run effect on adult outcomes, it may not be the result of persistence on any single skill impact in the short term, whether measured or unmeasured. This argument is not entirely new. In a review of the early intervention literature, Woodhead (1988) likened the unfolding of longer-term effects to a relay race. In contrast to the persistence of an initial intervention impact that transfers to adult functioning, a burst of performance in one skill domain could give rise to long-run effects through baton hand-offs to a second set of skills, despite fadeout of the first.

In line with Woodhead's analogy, one possibility is that interventions generate long-run effects by shifting individuals' trajectories in ways that promote the likelihood of subsequent educational inputs and experiences (e.g., *foot-in-the-door* processes; Bailey et al., 2017). For example, interventions that randomize children to slots at charter middle schools, like the KIPP program (Coen et al., 2019), strive to boost students' performance so that they may enter a strong high school and graduate better prepared for college. In the case of KIPP, persistent initial achievement effects are not a necessary precursor to long-run intervention impacts on educational attainment, so long as the initial boost opened the door to a subsequent beneficial context or experience.

In earlier work, we proposed a new theory that has some overlap with traditional skill-building explanations but attempts to reconcile fading impacts on targeted skills with longer-term impacts on adult outcomes: the *Large Interconnected Network Theory* (LINT; Bailey et al., 2024). Informed by the prevalence of fadeout across skill types, LINT depicts intervention impact fadeout as a regularity of development. LINT suggests that longer-term intervention effects could emerge as a product of small carry-over and transfer effects operating over time in large causal network of skills and environments (see Figure 2 in Bailey et al., 2024). While impacts on a given skill may diminish, an intervention could generate a network-level impact that sustains through small cross-over effects between impacted skills, contexts, and opportunities (Pages et al., 2023). Ultimately, impacts on adult functioning could be commensurate with the intervention's network-level effect in the medium-term (Figure 1, purple "Network-level effects (LINT)" line). 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 in the medium- to long-term[1], it may be difficult to make a priori predictions about how two interventions with equally sized impacts at posttest on two different skills will differentially affect long-run outcomes. The presence of dynamic "cross-over" processes among intervention impacts align with well-accepted theories about the many interrelated factors that shape development (e.g., Bronfenbrenner, 1992) and the complex systems that undergird psychological phenomena (e.g., van der Maas, 2024).

The Current Literature on Long-term Effects and the Need for Meta-Analysis

It is challenging to test causal skill-building mechanisms in the context of any single study. Indeed, it is not possible to identify the causal mechanisms driving long-run intervention effects by examining single point-in-time treatment estimates. Interventions are commonly "fat

[1] In the very long-term, impacts in a linear system 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.

handed" in their targets and content, plausibly capable of acting on a range of mechanisms (Eronen, 2020). A major challenge to reasoning about the causal mechanisms that systematically lead to longer-run effects across studies is that so few interventions collect adult follow-up data. The studies that follow individuals into adulthood are likely to form a non-representative sample that produced very large initial effects on many, or perhaps even one particularly compelling, outcome(s) (Watts et al., 2019).

Significant variation in *how* researchers compute and report impacts has made it very challenging to decipher the patterns of any one study and synthesize across studies. Indeed, researchers may selectively report main effects, change what skills they measure over time based on earlier patterns of results, report impacts on different subgroups across different papers[2], and draw conclusions about the intervention-effect "story" based on stand-out statistically significant findings.

If medium-term and/or adult outcomes are highly selected by researchers, for theoretical or other reasons, the same study might be consistent with more than one of the patterns presented in Figure 1. Take, for example, the patterns of effects observed in the Perry Preschool Program, one of the interventions that has exemplified the fadeout-emergence paradox. Anderson (2008) critiqued the inferences drawn from analyses of Perry data, pointing out the lack of consistency across models. Various analyses have ranged in sample size (i.e., from 60 to 120), differed in the treatment of gender as a moderator, and examined varying sets of outcomes. One long-run impact that has generated attention is the effect on adult criminal behaviors for boys, which has been argued to explain a large portion of Perry's public returns (Schweinhart, 2005). However, Anderson's re-analysis found weak support for impacts on crime for boys, and little evidence of

[2] See von Hippel & Schuetze (2025) for the risks of moderation without main effects.

replication in other early childhood evaluations. Anderson concluded that adult impacts for boys on selected crime outcomes were likely overstated, though narratives around Perry's long-run effects have nonetheless been shaped by this perceived effect.

Characterizing what happened in the short- and medium-term in Perry is also complicated. Posttest impacts on IQ—the primary outcome of focus at the time 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 story ("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 outcomes were more persistent in the medium-term (Heckman et al., 2013 Schweinhart, 2005) and impacts on a summary index of multiple adult outcomes were smaller than the likes of statistically significant effects on criminal activity for males (Anderson, 2008). Thus, depending on what outcomes researchers track over time, and how they synthesize evidence across outcomes, they may come to completely different conclusions about the mechanisms that produced the intervention's longitudinal effects.

Researchers' theories about the *specific* and testable pathways expected to underlie short- and long-run effects of programs are often not articulated and, when articulated, are commonly underspecified. When researchers do delineate specific mechanisms that they expected would drive their intervention's long-run effects, these theories are generally forgiving, fit a variety of observed results, and rarely generate clear falsification tests. When theories are more specific, it is often impossible to know whether they were formed *a priori* or *post hoc*. For these reasons, it is very challenging to decipher what lessons to draw from any one study's highlighted results. Looking across studies, it is even more difficult to try to gather a strong sense of which developmental theories find the most consistent support. Researchers' commonly form new and novel theories, thus distancing the relevance of any one study's findings from that of another and undermining the likelihood of making significant scientific advances, an issue Michel (2008) described as the "toothbrush problem" (wherein other researchers' theories can be likened to a strangers' toothbrush). Issues evident in the education intervention literature regarding selective reporting, lack of coordinated theory generation and confirmatory testing align with documented issues in psychology research writ large (e.g., Borsboom et al., 2021; Eronen & Bringmann, 2021; Fanelli et al., 2017; Ioannidis et al., 2017; Klein, 2014; Oberauer & Lewandowsky, 2019; Wagenmakers et al., 2012).

The Present Study: A Meta-Analytic Approach

Because of this field's focus on the specific and nuanced pattern of effects from individual studies, it is difficult to form priors regarding basic longitudinal trajectories of impacts across interventions. At present, it is challenging to know whether the fadeout-emergence pattern is broadly observed and whether short- and medium-term effects provide meaningful information about long-run impacts. For the field to make progress and check the extent to which basic theoretical assumptions hold, a transparent accounting of our current evidence base is needed.

In the present study, we attempted to gather all of the educational intervention RCTs implemented from infancy to middle school with impacts evaluated at the end-of-intervention and in adulthood. A sizeable body of causal evidence has accumulated across past decades, which enabled us to identify 29 unique treatment-control group contrasts from 25 studies.

Limiting the sample to RCTs made it possible to compute comparable, standardized main effects for each program that we could then analyze using meta-analytic techniques.

By examining dynamics among *average* intervention impacts using a diverse collection of studies, we hoped to overcome the "overfitting" problem of drawing strong theoretical conclusions from any one study. Our intention was not to test highly specific, intervention- and outcome-unique theories. Rather, we sought to capitalize on straightforward meta-analytic techniques to draw conclusions about core theoretical assumptions undergirding long-run intervention impacts. Indeed, at the heart of interventions targeting child skills, regardless of how diverse, is the causal expectation that changing early functioning will change later functioning.

Thus, we set out to examine two basic questions:

1. What is the pattern of longitudinal impacts observed across interventions? And, more specifically, is the fadeout-emergence paradox widely observed?
2. Do initial impacts on child skills predict long-run impacts on adult outcomes?

Method

Data

We created the MERF-Emerge dataset following many of the methods employed to create the original MERF sample (see Hart et al., 2024). Here, we provide a brief overview of the process of creating the dataset. The supplement contains additional details.

Inclusion Determinations

Figure 2 outlines the flow of reports and studies through our inclusion/exclusion criteria. We set out to create a meta-analytic dataset composed of all RCT-evaluated educational interventions in the United States that measured posttest and adult follow-up effects. In February

2024, the first author (a doctoral student at the time) conducted database searches through PsycINFO, ERIC, and EconLit which yielded 8,167 reports. Covidence identified 7,201 non-duplicate reports. At least one master's student-level research assistant and the first author then independently screened each unique abstract for initial inclusion. To be included at this stage, each abstract had to indicate that the study: (1) reported long-term follow-up impacts on adult outcomes, (2) was evaluated using RCT or lottery design, (3) was aimed at improving child or adolescent cognitive and/or social-emotional development (i.e., how we defined "educational intervention"), and (4) took place in the United States. Peer-reviewed manuscripts, working papers, preprints, and government reports were included.[3] We reached consensus on discrepancies, and together identified that 155 of the 7,201 abstracts passed abstract screening.

At least one research assistant and the first author then conducted a full-text review of the 155 reports that passed the initial review. At this stage, the first author reviewed the reports in more detail to confirm that the included studies met the inclusion criteria. We added two dimensions to our criteria at this stage: to ensure that the included studies showed strong construct validity as reporting intervention impacts on *adult*, *long-run* follow-up effects, we only included interventions that (1) ended prior to students' entry to high school and (2) that contained follow-up assessments after age 18 or grade 12. Eighty reports from 28 unique studies passed the full-text review. At this stage, we included four additional reports from four studies that we knew met our inclusion criteria but were not identified through our search process.[4] Thus, while we have confidence that we identified the vast majority of studies that met our criteria, that we did

[3] 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.

[4] Three studies were interventions that we knew would meet our inclusion criteria that were included in our original MERF meta-analytic sample: 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 proceeding forwards/backwards search process.

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.

We then conducted a rigorous search process to gather all the papers reporting intervention effects for these 32 studies. For each of the 84 reports, two research assistants independently conducted a series of "forward" searches to identify reports containing follow-up effects and "backward" searches to identify reports that contained posttest or medium-term follow-up impacts. The first author and the fourth author (a Masters-student-level research assistant at the time) worked together, in consultation with the two doctorate-level investigators on the project (second and fifth authors), to make final inclusion determinations and to identify what reports to code. We identified 25 studies that met the inclusion criteria and reported posttest effects on at least one outcome.

For these 25 studies, our search produced 882 reports which we then reviewed for inclusion in the coding phase. When identifying reports 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 six 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 reports 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 reports that provided estimates on the same outcome, we chose to code the report 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). We coded a total of 94 unique reports (101 when summing the number of individual reports for each study, not eliminating reports that reported impacts for multiple studies). The 25 studies reported impacts for 29 treatment-control-group contrasts generated through random assignment (henceforth referred to as "interventions").

Table S1 provides in-depth details on each intervention included in our sample, including key program characteristics, targets, and notable details regarding the coding process and inclusion criteria adherence. Overall, the sample of 29 interventions was challenging to code given the number of papers reporting impacts on each intervention and unclear reporting within and across papers. Importantly, there were some studies for which we identified violations to our inclusion criteria. In many cases, these problematic study details were inconsistently detailed across reports. For example, 2 interventions provided booster sessions in high school, 1 study excluded 46 "non-white" participants from the sample, 2 interventions reported posttest impacts collected prior to intervention end, and 2 interventions reported posttest impacts measured with a notable delay after intervention end. These deviations are detailed in Table S1. We ran various sensitivity analyses to probe these study-specific issues, including leave-one-out models, and generally found that our results were robust.

Coding

The first author and third author (also a doctoral student) then double coded the 94 reports. Prior to double-coding each paper, we reached 79% reliability on a random sample of 5 reports from the 80 that were identified following full-text review. Our reliability across all reports was 89%. An independent research assistant identified discrepancies, which the coders then discussed. Consensus was established in consultation with the doctorate-level authors as needed. For each study, we coded a broad array of information regarding the treatment characteristics and study features. We coded posttest, six-month to two-year medium-term follow-up, and adult follow-up impacts, as well as information relevant to each impact (e.g., when the data was collected, the measure that was used, descriptive data).

After coding, we compiled a final analytic sample of posttest, medium-term, and adult follow-up impacts. At this stage, we dropped outcomes that provided insufficient details necessary for computing effects. This included cases in which inadequate information was provided to compute effects or to determine the valence of the outcome (i.e., is a higher score "better" or "worse") and, thus, it was not possible to properly scale the effect.

Effect Size and Standard Error Estimation

We then took steps to compute effect sizes and standard errors for each outcome in our dataset. Insofar as the study authors reported effect sizes in standardized units, we opted to use the researcher-reported statistics. However, in many cases, intervention effects were either reported in units other than standard deviations or were altogether unreported. In these cases, we used various approaches to estimate treatment impacts depending on the information that the authors provided. Supplemental Figure S1 details our approach. Of note, we only included main treatment effects in our meta-analysis. When impacts were reported for subgroups, we backed out main effects by computing a weighted average of subgroup impacts and standard errors, so long as the necessary information was provided.

When possible, we converted the author-reported statistics (e.g., odds ratio, predicted change in probability, hazard ratio) to estimated Cohen's-d-like effects in standard deviation units. When only descriptive data (i.e., means, proportions) were reported, we used this information to back out 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. We rescaled all outcomes to ensure that a larger effect size indicated a preferable outcome for those in the treatment group.

Other than in the few cases when the study authors 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; pg. 27):

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

Of note, we ran supplemental models that were weighted by other precision estimates including sample size, and an alternate set of standard errors generated through more complex outcome-specific formulas (see Figure S2) incorporating model-based parameters (e.g., confidence intervals, p-values) and adjustments for dichotomous outcomes. Our substantive conclusions were similar to those gleaned using standard errors computed through our primary formula, presented above.

Several studies in our sample used cluster-based randomization. To appropriately 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 based on a model containing cluster adjustments.

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. Medium-term follow-up included 6-month to 2-year follow-up assessments. Following Hart et al. (2024), medium-term follow-up assessments were split into two bins for analysis: 6- to 12-month follow-up and 1- to 2-year follow-up. On average, studies measured adult effects four times throughout adulthood, though as few as one time and as many as eight times. Figure S3 depicts the assessment schedule for each study. Our models considered adult impacts simultaneously, regardless of assessment timing. In supplemental analyses, we probed how assessment timing affected our results. We did not find evidence that assessment timing meaningfully impacted our results.

We estimated the meta-analytic average of intervention impacts at each assessment wave (i.e., posttest, 6- to 12-month follow-up and 1- to 2-year follow-up, and adult follow-up) in R. First, we used the *metafor* package to compute meta-analytic estimates using inverse variance weighting ($\frac{1}{se^2}$), and a random effect for study (Viechtbauer, 2010). Next, we used the *clubSandwich* package to estimate cluster-robust standard errors (clustered at the study level) through robust variance estimation (RVE), with default (i.e., “CR2”) small sample size adjustments (Hedges, Tipton, & Johson, 2010; Pustejovsky, 2023; Pustejovsky & Tipton, 2018; Tipton, 2015). 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 (“aligned groups” as described in the following section).[5]

[5] 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

Medium-term Persistence

We next investigated the extent to which posttest impacts on child skills persisted in the 6 months to 2 years following the end of the interventions. We used methods described in depth in Hart et al. (2024). In short, we first identified "aligned groups" in the data: cases in which the same construct was measured using the same instrument, reporter, and subscale at posttest and at least one follow-up within the same intervention. By analyzing impact trajectories among aligned groups, we were able to move beyond simply examining average impacts at each assessment wave to identify how effects on the same construct and measure changed across time. Limiting to the same outcomes increased the validity of our estimates as measures of "fadeout" and increased our confidence that observed changes in intervention impacts were due to the time elapsed between the intervention and follow-up, rather than changes in the measure.

In two separate models, we regressed impacts collected at 6- to 12-month follow-up and 1- to 2-year follow-up (i.e. medium-term follow-up) on posttest impacts on the same skill.[6] These models produced two theoretically meaningful terms. First was the slope term, or "conditional persistence rate," which indicated the extent to which posttest impacts on a particular skill persisted at follow-up, conditional on the magnitude of the posttest effect (i.e., a slope of 1 would indicate 100% persistence of the posttest impact). Second, the intercept term captured the portion of the follow-up effect that was unexplained by posttest impacts on the same skill (i.e., unmeasured mediational effects). We estimated the models as follows:

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.

[6] In the case that there was 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.

Equation 1

Level 1:

$$ES_{ifg} = \beta_{0s} + \beta_1 ES_{ipg} + \varepsilon_{sfg} \tag{2}$$

Level 2:

$$\beta_{0s} = \gamma_{00} + u_{0s} \tag{3}$$

where *i* represented intervention, *s* represented study, *f* indicated the follow-up assessment wave (6- to 12-months or 1- to 2-years), p indicated the posttest assessment wave, and *g* indicated the aligned grouping of construct and instrument. As such, ES_{ifg} represented follow-up impacts on the same skill as measured at posttest (ES_{ipg}). Thus, β_1 indicated the conditional persistence of posttest impacts at follow-up and β_{0s} indicated the study-level average portion of the follow-up impact that was unexplained by posttest impacts on the same skill (plus the unique random effect for each study, u_{0s}).

We weighted the models by the inverse sampling variances of the follow-up effects, entered a random intercept for study, and used cluster-robust standard errors (clustered at the study level). 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.

Correspondence between Posttest Impacts and Adult Follow-Up Impacts

Approach. To examine the association between posttest impacts and adult impacts, we used an adapted model of *Equation 1.* Here, we were forced to rely on average impacts across outcomes, because there was no clear 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, our primary model regressed intervention-level average adult follow-up impacts on intervention-level average posttest impacts. 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.

Analytic Models. Regressing average adult follow-up impacts on average child posttest impacts produced a slope term and an intercept term that were both theoretically relevant to our research question. These terms were in some ways similar to those that we estimated when we examined the persistence of short-term impacts (i.e., *Equation 1*), but different in that adult follow-up measures were not aligned with child posttest measures. Hence, the slope term from the model estimated the effect of initial posttest intervention impacts on adult follow-up intervention impacts. Of note, it may be assumed that the term also captured bias due to omitted variables (i.e., the effect of intervention impacts on other unmeasured factors correlated with posttest impacts). The intercept represented the portion of the adult intervention effects that was not explained by posttest impacts. Thus, the intercept captured intervention impacts at follow-up that must have been driven by mechanisms not captured by, and uncorrelated with, posttest impacts on measured child skills. If interventions were “black boxes,” then we would expect to find a large intercept term and small slope term (Pages et al., 2023). 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 smaller intercept term and a larger slope term. Both the predictor and outcome were measured with measurement error. Insofar as the measurement error was uncorrelated, then we would expect to find a larger intercept effect. Insofar as the error was correlated, then the slope should be considered an upward bound for the true association between child and adult impacts.

Results

The Sample of Educational RCTs that have Measured Child and Adult Effects

As detailed in Table S1, the sample comprised 25 studies and 29 unique interventions groups (i.e., treatment-control contrasts). The average implementation year was 1991, though programs started as early as 1963 and as late as 2013. The average posttest sample size was 916, though sample sizes ranged from 33 children to 10,170 children. Interventions lasted about 2 years on average. Program duration ranged from 1 month to 8 years. Interventions started as early as during mothers' pregnancy and, by design, as late as at the end of middle school (i.e., ~ 14 years old). On average, children were about 7 years old at intervention start. Many of the interventions in our sample were broad in scope: 72% targeted parents, 93% targeted social-emotional skills, and 48% also targeted cognitive skills. Only 1 intervention targeted only cognitive skills.

Table 1 details the composition of the posttest and adult impacts in our sample. Our 29 interventions reported 179 posttest impacts. On average, each intervention reported posttest impacts on 6 outcomes, though this ranged from 1 to 25 outcomes. 74% of the posttest impacts were estimated using social-emotional measures and 26% of impacts were estimated using cognitive assessments. The meta-analytic average post-test impact was .24 *SD* (p = .006) for cognitive outcomes, and .08 *SD* (p = .06) for social-emotional outcomes.

Impacts were collected for more outcomes at adult follow-up (n = 497) than at posttest, though many studies measured adult effects at multiple assessment waves (average = 2.72 adult assessment waves per intervention; range = 1 to 8). On average, adult outcomes were measured about 14 years after posttest, though as early as 4 years after posttest and as late as 49 years after posttest (see Figure S3 for intervention-specific assessment timelines).[7] The sample contained a diverse collection of adult outcomes including outcomes related to substance use (28%), education (23% of outcomes), psychological wellbeing (14%), employment (10%), cognitive functioning (6%), crime (6%), health (4%), and social service receipt (1%). Outcomes that did not fit these categorizations (10%) comprised the "other" category. Interestingly, impacts on these diverse outcomes were generally around the magnitude of .10 *SD* to .20 *SD*, though average impacts for each category ranged from .04 *SD* to .28 *SD*.

What Kinds of Programs are in the Sample?

A diverse collection of educational interventions has achieved the feat of assessing impacts into adulthood. Here, we attempt to broadly categorize the programs included in our sample. Table S1 provides additional information on each intervention, including references to papers reporting results.

Early Care and Education. The first category of studies comprises early care and education programs. Interventions targeting young children were the first RCT-evaluated educational programs that measured impacts into adulthood in the United States. Programs were first implemented in the 1960s, with most evaluations concentrated from then until the 1980s. One set of programs examined the impact of early classroom-based care (versus no care) for preschool-aged children, often coupled with additional family and health services. These included the extraordinarily influential Perry and Abecedarian programs, in addition to the Early Training Project and Project Care. The Chicago School Readiness Project examined the effects of a specific curriculum in Head Start centers (versus Head Start as usual), marking the only

[7] 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.

early-childhood-focused curricular intervention that has measured long-run intervention impacts, to our knowledge. A second set of programs—the Infant Health and Development Program, and two Nurse Home Visiting evaluations—targeted younger children, and centered around home visiting (with center-based care provided in some cases).

Elementary-Level. In the elementary years, there were four curricular intervention evaluations that measured effects into adulthood. Two interventions were focused on the development of academic skills, one involving reading remediation for struggling readers, the other an examination of special education instructional approaches. Two other studies evaluated the effects of the Good Behavior Game, a universal classroom-based behavioral management program.

Additionally, the Project Star experiment focused on elementary school class size while the Multimodal Treatment Study of ADHD entailed a multi-condition evaluation of behavioral and pharmacological treatments, of which we focused on the behavioral arm.

Middle School Transitions. Three programs focused on middle school as a critical transition period that shapes long-run success. Two involved lottery evaluations of charter middle schools in which children were randomized to receive or not receive a slot, including the Knowledge Is Power Program (KIPP) middle schools, and a national synthesis of 33 charter middle school lotteries. A third program, Higher Achievement, involved intensive afterschool and summer programming for precocious middle schoolers, aimed at preparing students to successfully apply to competitive high schools.

Middle School Positive Youth Development. The final programmatic category included positive youth development programs aimed to promote prosociality and prevent problematic behaviors (e.g., substance use) in middle-school-aged children. The programs in this category largely focused on supporting children to develop coping strategies and life skills that would enable them to avoid risky behaviors in adolescence and beyond. Often these programs also targeted parents with the goal of strengthening parent-child relationships and equipping caregivers with skills for managing their child's behaviors. 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 additional programs were focused on specific populations, with intervention content tailored accordingly: New Beginnings, for children with divorced parents, and Strong African American Families, for Black students.

Descriptive Patterns of Longitudinal Intervention Impacts

We first examined the descriptive patterns of intervention impacts across assessment waves to gain a broad sense of the trajectory of effects (see Table 2 and Figure 3). Overall, we observed that interventions generated initial effects, that these effects faded, and that programs nonetheless produced long-run adult impacts. However, the observed pattern did not follow quintessential portrayals of the fadeout-emergence paradox (see orange "Fadeout-Emergence" in Figure 1). Instead, long-run adult effects appeared to be similar in magnitude to faded out medium-term effects.

Specifically, across all outcomes and interventions, we observed a meta-analytic average 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; see Table 2, Panel A). We observed considerable heterogeneity in effects across studies, τ = .24 at posttest and τ = .11 - .15 across the follow-up waves. The 95% prediction interval (i.e., inferred distribution of true underlying effects) ranged from negative to positive at all assessment waves.

Figure S4 visualizes all of the effect sizes at each assessment wave, highlighting the heterogeneity in estimates.

We also examined these patterns using a sample of effect sizes collapsed at the intervention level, in preparation for subsequent analyses in which we predicted average adult impacts using average posttest impacts. With this method, 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). Aggregated at this level, we observed less between-study heterogeneity and narrower prediction intervals.

Policy researchers understandably present longer-term impacts in raw (unstandardized) units; for example, an impact of 3 percentage points (pp.) on college enrollment may be more interpretable to most readers than an impact of .05 SD on college enrollment. However, some potential explanations of the fadeout-emergence paradox make specific predictions about the relative magnitudes of impacts on different kinds of outcomes. For example, consider a hypothetical scenario where impacts on student test scores fade from .20 SD to a statistically nonsignificant .02 SD in the years following the intervention, and the impact on college enrollment is a statistically significant 3 pp. (.05 SD). The reader may infer that this is an example of fadeout-emergence. However, when the long-term impact is also provided in SD units, it becomes clear that it is well within the confidence interval of the medium-term impact, and is entirely consistent with a pattern of findings where medium-term and long-term impacts are randomly distributed around the same mean effect size. To provide more context to the SD-scaled impacts presented here, a .05 *SD* impact in our sample was equivalent to an approximately 3 pp. increase in college enrollment, and 1 point increase on a cognitive assessment.

Persistence in the Medium-term

To further examine the rate of posttest persistence in the years after programs ended, we modeled the proportion of the initial effect that persisted at both medium-term follow-up assessment waves. Before doing so, we first checked the descriptive trajectories of impacts on the *same outcomes* measured consistently at posttest and medium-term follow-up (see Table 2, Panel C). Considerably fewer interventions (k = 13) reported aligned impacts for at least one outcome. We observed a larger posttest effect of .20 *SD* (*p* = .09) among outcomes that were measured consistently, that then 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. For interventions that assessed the same outcomes, we observed an adult impact of .10 *SD* (*p* = .15), closely matching what we observed in the larger sample.

We then estimated what portion of posttest impacts persisted at medium-term follow-up by regressing follow-up effects on posttest effects (Table S2). At 6- to 12-month follow-up, we observed a conditional persistence rate of .51 (p = .05) and a small intercept effect of .05 (*p* = .11), suggesting that conditional on the posttest impact magnitude, effects persisted at a rate of about 50%. At 1- to 2-year follow-up, we observed a reduced conditional persistence rate of 26% (*p* = .10) and an intercept of .04 (*p* = .56).

Forecasting Adult Follow-up Impacts using Posttest Impacts

We next turned to testing whether intervention impacts on child skills measured at intervention end forecasted impacts measured in adulthood. We used average impacts and, as such, each intervention contributed one posttest impact and one adult follow-up impact to our model. Figure 4 plots the average posttest and adult impact for each study (see Table S3, column 1). As demonstrated by the magnification of the origin in the upper-right corner of the figure, several studies produced near-zero average posttest effects and adult follow-up effects. Most interventions generated consistently positive impacts that fell within the upper-right quadrant of Figure 4. However, some studies produced posttest and/or follow-up effects that were negative on average, as represented in the three other quadrants.

Our meta-regression model confirmed a statistically non-significant and imprecise association among posttest and adult follow-up impacts of β_1 = .13 (*p* = .30). The slope of .13 indicated that, on average, studies that generated larger posttest effects tended to also generate larger follow-up effects, though the association was far from a 1:1 correspondence. Put differently, 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 effect of .03 *SD* (*p* = .09) indicating that a considerable portion of adult impacts (.03/.04 or 75%) was explained by factors other than the intervention impacts on child skills at posttest.

In supplemental analyses that dropped each study (see Table S4), we observed that the modeled slope was significantly influenced by the Perry Preschool Intervention. Perry produced the largest average posttest effect size and largest average adult effect size in our sample. As depicted in Figure 4, when we dropped the intervention's estimates from our model, the slope reduced from .13 to .04 (*p* = .68). It is also worth noting that the Higher Achievement intervention was another influential data point; upon removing Higher Achievement, our slope increased from .13 to .21 (*p* = .09).

Publication Bias and Selective Reporting

We took several steps to examine the extent to which our estimates were biased by selective reporting and publication issues. One might be concerned that interventions in our

MERF-Emerge sample were substantively different from interventions in the literature that have not collected adult follow-up data. Interventions that assess adult follow-up outcomes may have shown particularly promising posttest or medium-term effects that generated the enthusiasm and funding necessary for long-run data collection (see Watts et al., 2019 for discussion).

However, we did not find evidence that posttest impacts for studies in our sample were larger than that of other educational interventions. If anything, the posttest impacts in the MERF-Emerge sample were smaller than those observed in the larger MERF sample (.15 *SD* on average at the outcome level in this sample versus .20 *SD* in MERF; Hart et al., 2024). In Figure S5, we plotted the forecasted long-run adult impacts for studies in the MERF sample that did not measure adult outcomes. To do so, we used the studies' posttest impacts and the estimated associations among posttest and adult impacts from the present study (.13). The figure depicts minor differences in the forecasted long-run impacts of studies that did not measure adult effects, highlighting that posttest impacts in the MERF-E sample were not very different from those observed in the larger literature.

Relevant to selective data collection and/or reporting, the average posttest impact for outcomes that researchers consistently measured at medium-term follow-up was larger than the average posttest impact for all outcomes in the sample (posttest effect of .20 *SD* among aligned outcomes versus .15 *SD* without alignment considerations; see Table 2, Panel C), suggesting the possibility of bias in what outcomes were measured at follow-up. We observed a similar pattern in the original MERF sample (posttest effect of .29 *SD* among aligned outcomes versus .20 *SD* without alignment considerations). Conditional persistence rates were very similar, albeit slightly stronger, in this sample compared to the larger MERF sample at 6- to 12-months follow-up (51%

in this sample versus 45% in MERF) and 1- to 2-years (26% in this sample versus 18% in MERF).

To further examine issues of publication bias and selective reporting in our sample, we generated funnel plots (see Figure S6) and examined the association between estimate precision and estimate magnitude using PET-PEESE models (Chen & Pustejovsky, 2025; Stanley & Doucouliagos, 2014). The funnel plots suggested the possibility of a slight positive skew wherein larger positive effects were reported more than expected. Our PET test indicated that studies with less precise estimates observed larger effects. Indeed, less precise (i.e., larger) standard errors were predictive of larger effects 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). While it may be the case that smaller-sample studies reported more selected impacts, it may also be the case that smaller-sample studies generated larger effects through more intensive programming that is otherwise challenging at scale.

In line with the recommendations of Stanley and Doucouliagos (2014), we examined the PET-adjusted estimates to gain a sense of the expected meta-analytic average for studies with perfect precision (i.e., standard errors = 0). As presented in Table S6, our meta-analytic averages reduced by ~40 to 90% when controlling for precision. However, the PET model is known to underestimate non-zero effects and provides a crude estimate considering the substantive reasons why smaller studies may produce larger effects.

We then examined the distribution of statistically significant p-values using the approach developed by Simonsohn and colleagues (2015). While there has been recent debate regarding the validity of this method for detecting bias under all conditions (see McShane et al., 2016; Morey et al., 2025; Simonsohn, 2025), we believe that this method nonetheless provides an interesting descriptive profile of p-values in our sample. The p-curve analysis did not suggest any major upticks in the number of p-values near the .05 conventional "statistical significance" cut-off. The analysis did, however, indicate that power to detect statistically significant estimates was particularly compromised at the medium-term follow-up assessment waves. That we did not find an uptick of p-values near .05 may in part reflect our coding approach which prioritized computing effects using descriptive information (i.e., estimates from correlation matrices or descriptive tables) when standardized impact estimates were not provided. Of note, we examined the funnel plots and p-curve models using alternative standard errors, computed through more complex methods, and most notably found that the p-curve for adult impacts estimated lower power for detecting real effects (see Table S5, Figures S8 and S9).

We additionally examined two selection models to further probe the extent to which our meta-analytic estimates were biased by selective reporting of more statistically significant results (see Table S6). Using the techniques detailed by Vevea and Woods (2005), we found that after accounting for the likelihood that more statistically significant estimates were reported more consistently, our estimates were estimated to be inflated by 0 *SD* to .03 *SD* across assessment waves. We then examined the estimates from a more recent model proposed by Pustejovsky, Joshi, and Citowicz (2025) which accounts for the nested and non-independent nature of our data. We set two "steps" at p = .025 and p = .50. Interestingly, the model suggested that while the medium-term follow-up impacts were likely *over*-estimated by ~80% (i.e., considerably more than that found using the Vevea & Woods approach), posttest and adult follow-up effects were likely *under*estimated by ~40 to 50%. To examine whether the results were consistent with a reasonable alternative model with more steps, we ran the model again with nine steps at p = .025, .05, .10, .20, .30, .40, .50, .70, .90. Results were not consistent with the two-step model; most notably, whereas in the two-step model adult follow-up effects were underestimated, in the nine-step model adult effects were judged to be overestimated by about 50%. Although no one model provided definitive results, together, the models suggest that there does not appear to be egregious evidence of selective reporting and publication bias in our sample in the direction we would expect (i.e., underreporting of small and statistically non-significant results).

Additional Analyses

Next, we performed a series of exploratory analyses and robustness checks to further examine the associations among intervention impacts on child and adult outcomes and the sensitivity of our primary estimates.

Exploratory Analyses

LINT Simulation. Given the striking similarities between the observed longitudinal effects in our sample (see Figure 3) and our theoretical depiction of “Network-Level Effects” (see Figure 1), we examined whether the observed pattern aligned with simulated estimates based on the LINT theory (Bailey et al., 2024). Recall that LINT proposed that intervention impacts on various outcomes in a network interact to generate adult effects despite skill-specific fadeout. Figure S10 depicts the trajectory of effects for aligned outcomes in the medium-term, linked with intervention-level average adult impacts, as well as the simulation-based estimates (more details provided in the supplement). The simulated effects largely followed the patterns observed in our data.

Differences by Outcome Type. We then tested the extent to which the observed associations varied by posttest and adult outcome types (see Table S7). 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. Overall, we observed stronger correspondence between cognitive posttest impacts and adult effects, as compared to that for social-emotional posttest impacts, though the difference was not statistically detectable (see Figure S11). In fact, the association between social-emotional posttest impacts and average adult effects was negative.

Next, we examined the correspondence between average posttest impacts and adult outcomes aggregated by the adult outcome categories with the most data coverage. The correspondence between posttest and adult effects was smaller for educational outcomes than for other outcomes, and the slope for substance use outcomes was larger than for other outcomes. It is possible that there was greater alignment in posttest and adult assessments of substance use than other adult outcomes given that substance prevention interventions in our sample measured similar substance-related outcomes at posttest and adult follow-up.

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 S8). All the differences were not statistically significant, and the sample sizes for these models were limited. In line with Table S7, point estimates trended towards cognitive posttest impacts as being more predictive of adult educational and psychological wellbeing outcomes than social-emotional impacts. In contrast, social-emotional posttest impacts were more predictive of substance-related outcomes than cognitive posttest impacts.

High-Occurrence Adult Outcome: High School Attainment. Next, we examined the association between average posttest impacts and the adult outcome that was reported most consistently across interventions: high school attainment (see Table S9 and Figure S12). Our sample contained 12 interventions that reported impacts on high school attainment.[8] Overall, we found much stronger correspondence between posttest impacts and high school attainment (β = .36) than we observed for all adult impacts, though this estimate was not statistically significant (p = .10).

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. The results from these models are presented in Table S3 and Figure S13. When we examined the associations between medium-term follow-up impacts and adult effects using all reported impacts, we found that medium-term impacts were more predictive of adult outcomes than posttest impacts were. Six- to 12-month follow-up impacts predicted adult follow-up effects at a rate of .27 (p = .07) and 1- to 2-year impacts predicted adult impacts at a rate of .94 (p = .02), suggesting very high correspondence between medium-term follow-up effects and adult effects. However, these estimates came from a limited number of interventions that reported medium-term follow-up. 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.

We then 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

[8] In the case that an intervention reported impacts on 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.

differences in the associations with adult follow-up effects using the same sample. While we observed a near-zero estimate for the association between 6- to 12-month impacts and adult impacts using aligned outcomes, the association between outcomes that continued to be measured at 1- to 2-year follow-up and adult effects remained larger (β = .71; *p* = .15) than the association observed for posttest impacts. These patterns suggested more support for the possibility of 1- to 2-year follow-up effects as consistently predictive of adult effects, though the standard errors on all estimates were large due to the very small sample size (only 8 interventions), making it challenging to draw firm conclusions.

Adult Assessment Timing. Finally, we explored whether posttest impacts on child skills were better forecasters of more proximal adult impacts (see Table S10). Indeed, there was a great deal of variation in when adult effects were assessed, from when participants just entered adulthood to middle age. The contribution of age at adult assessment was not statistically significant. Contrary to our expectations, the interaction between posttest effect size and age at adult assessment was positive (β = .01; *p* = .47), suggesting that, if anything, the association between posttest and adult impacts was larger at further out adult follow-up waves.

Robustness Checks

We also ran a series of robustness checks to further examine the robustness of our meta-analytic averages and estimates of the correspondence between posttest and adult effects. We briefly summarize the models here, with additional details provided in the supplemental text. First, we ran models where we dropped each study to examine how study-level idiosyncrasies (such as those described in Table S1) may have biased our results. For 23 of the 25 studies, dropping the study-specific estimate minimally affected the slope and intercept estimates (see Table S4). However, as mentioned earlier, dropping Perry Preschool led to a notably reduced slope and dropping Higher Achievement led to a notably increased slope. Thus, the magnitude of the slope estimate appeared sensitive to the inclusion of these two interventions.

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 (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. As shown in Table S5, the results using alternative weighting and CHE were substantively aligned with our main findings. The CHE model demonstrated that there was considerable within-study heterogeneity in the meta-analytic estimate of intervention effects at each assessment wave.[9]

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. Accounting for average posttest impact did not explain much of the within-study variation ($\tau = .21$ to $\tau = .20$). The association between intervention-level posttest impacts and outcome-level adult impacts using this specification was $\beta = .32$ ($p = .02$). 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.

[9] 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).

Building from our exploratory analysis examining assessment timing age (Table S10), we investigated whether our results were robust to assessment-timing-related issues. Upon controlling for participant age at adult follow-up, time elapsed between posttest and adult assessments, and child age at the time of the intervention the estimates were similar to those from our primary models (Table S11).

Additionally, we tested whether several study- and data-related irregularities may have biased our results (see Table S12). Additional motivation is provided in the supplement, but, in short, we ran models in which we dropped: a) high-attrition adult outcomes, b) studies that collected posttest measures before the end of the treatment, c) "lifetime" or "ever" measures, and d) effects whose estimation was more computationally heavy and/or relied on greater assumptions (e.g., the transformation of odds ratio effect sizes to Cohen's-d-like effect sizes). We also tested an alternative approach to estimating intervention-level average adult effects in which we first aggregated effects at the outcome domain level, to attenuate the influence of many effects reported for the same outcome (see supplement for additional discussion). All these models produced estimates in line with our primary results with one exception: dropping effects that required a great deal of estimation, which included all dichotomous effects, increased the correspondence between posttest and adult effects considerably (β_1 = .61, p = .06). However, upon visual inspection of the distribution, the Perry Preschool program appeared as a potentially highly influential data point, as it was in the full model (i.e., large posttest effect and large adult effect). Upon removing Perry Preschool from the model, the slope attenuated considerably (β_1 = .37, p = .14), though was still larger than our primary estimate.

Discussion

The mystery of sleeper effects—whether they're real and what they imply—has long captured the interest of social scientists. 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 et al., 2009; Gray-Lobe et al., 2023). These patterns, coupled with increasing evidence for the ubiquity of medium-term fadeout (Hart et al., 2024), have left researchers to grapple with several questions regarding how childhood interventions affect life course outcomes, if not through persistent impacts on targeted child skills.

Motivated by these unresolved questions, the current study set out to conduct a broad examination of the longitudinal impacts of RCT-evaluated educational interventions. We first evaluated the extent to which the fadeout-emergence pattern was widely observed. Of course, we were limited to examining the sample of studies that have followed participants into adulthood, which is likely a highly selected sample that does not reflect all educational RCTs (Watts et al., 2019). We identified 29 interventions targeting children prior to high school that have measured initial and adult effects.

Among these 29 interventions, we did not find widespread evidence of *emergent* or *sleeper* effects as they have been most commonly construed (i.e., a U-shaped curve for which the fadeout occurs in the medium-term, followed by a sharp resurgence of adult effects; see Figure 1). Instead, we found that initial effects on child skills faded considerably in the two years after programs ended. Nonetheless, programs generated statistically significant impacts on adult outcomes, and the magnitude of these effects (.04 to .11 *SD* depending on modeling approach) was very similar to the magnitude of medium-term, faded-out impacts on child skills. As such, longitudinal program effects appeared to most closely align with the hypotheses of LINT (Bailey et al., 2024), reflected by the "Network-level Effects" pattern in Figure 1.

A number of important features of our methodological approach likely shaped our observation that long-run effects did not align with typical conceptualizations of emergent or sleeper effects. First, we went to great lengths to compute intervention impacts on all posttest and adult outcomes, including those that researchers did not highlight or even explicitly report (e.g., effects that we computed using information provided in correlation matrices and descriptive tables). Importantly, we only analyzed main impact estimates (in some cases recovered from subgroup estimates) rather than subgroup estimates, which can suffer from reporting and selection bias (von Hippel & Schuetze, 2025). We also examined intervention-level averages that aggregated effects across all outcomes, rather than focusing on individual outcomes. Our approach allowed us to gain a broad understanding of impacts across all the outcomes that were sufficiently theoretically aligned to have been assessed. In doing so, our approach de-emphasized stand-out intervention impacts that researchers often highlight, effects that have likely fueled the expectation of a U-shaped pattern of large emergent adult impacts.

After establishing that programs can generate long-run effects despite medium-term fadeout, we then probed the question of *how*. A basic assumption undergirding program-specific theories is that interventions that have larger impacts on child skills at program end should also generate larger effects on adult outcomes. Thus, we examined whether there was any correspondence between initial intervention impacts on child skills and long-run impacts on adult outcomes. In line with our descriptive observations, we observed strong correspondence between medium-term follow-up impacts and long-run adult effects among the subsample of studies that reported medium-term follow-up. However, we did not observe consistent evidence that average initial impacts on child skills were associated with average long-run impacts on adult outcomes. We observed a positive (non-statistically significant) association between average posttest impacts on child skills and average follow-up impacts in adulthood of $\beta_1 = .13$, but we also observed a .03 *SD* intercept effect, suggesting that 75% of the average .04 *SD* adult impact was explained by factors not captured by posttest assessments of children's skills. The slope also appeared sensitive to the inclusion of several specific studies. Removing Perry Preschool reduced the slope to .04 and removing Higher Achievement increased the slope to .21.

Further exploratory analyses suggested that the association among child skills and adult effects can additionally vary depending on how child skills and adult outcomes are operationalized. In one exploratory analysis, we found that posttest impacts on cognitive skills, as opposed to social-emotional skills, were more predictive of long-run outcomes (β_1 = .16 versus β_1 = -.06). In a second analysis, we observed greater correspondence between average posttest impacts and impacts on the adult outcome measured most consistently across studies: high school graduation (β_1 = .36).

Together with the finding of medium-term fadeout, the association between posttest and adult effects suggests that it is highly unlikely that persistence on a well-defined, small number of "key" child skills—targeted, and impacted, by interventions—is likely to fully explain the long-run effects of programs. Skills may beget skills, but in far more complicated ways, commensurate 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.

Importantly, we do not view these results as cause for despair regarding the state of our field and intervention science. Imagine a hypothetical scenario in which 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 indicates the causal long-run impact of changes to child skills targeted by interventions. Although .10 is far from 1:1 correspondence between targeted child skills and important adult outcomes, .10 indicates some correspondence between the short-run success of an intervention and the longer-term implications of that intervention. Although the current data does not allow us to make precise predictions about which interventions will reliably generate long-run effects, the slope of .10 does indicate that post-test impacts on child skills may contain important information when making investment decisions based on longer-term projections.

Certainly, our reliance on average impacts across studies may leave readers who hold specific theories about intervention mechanisms and skill development disappointed. We agree that it must sometimes be the case that a particular skill relates to, and influences, the development of some skills more than others. However, in line with the skill-type-null hypothesis, and in the context of forecasting long-run adult effects, we anticipate that it is very challenging to identify *a priori* reliable skill-to-skill combinations given the highly complex mechanisms that undergird the production of long-run adult effects (i.e., LINT-like processes). However, researchers should certainly continue to interrogate whether there are specific child skills that causally affect adult outcomes of interest.

We were surprised to observe much stronger correspondence between average medium-term follow-up impacts and average adult follow-up effects—ranging from 0 to .94 depending on our approach (average β_1 = .42)—although our estimates relied on a small subsample of interventions and were generally imprecise. Nonetheless, the associations indicated some promise for the strategy of using average medium-term follow-up effects to proxy the network-level impact of programs, which may correspond strongly with adult effects. More research is necessary to identify exactly when, if at all, medium-term follow-ups become predictive of adult impacts.

Commensurate with relatively strong associations among medium-term and long-run effects, the trajectory of longitudinal effects that we observed appeared to most closely align with the hypotheses of the Large Interconnected Network Theory (LINT; Bailey et al., 2024). While the current study did not explicitly set out to test LINT, our findings appear to resemble the predictions of LINT (as probed using simulations in Figure S10). LINT argues that the effects of an intervention on any one skill are likely to fade, but, nonetheless, ripple out to a network of other child skills, relational factors, and contextual factors (Bailey et al., 2024). Through interactive, cross-lagged associations among all impacted domains, interventions could generate a small, but stable, network-level impact on functioning that is sustained through ongoing cross-lagged processes. This network-level impact is expected to then extend to adult functioning, generating adult effects of a similar magnitude as impacts on child skills observed in the larger network. Our findings warrant future *a priori* examinations of the theory.

Limitations and Recommendations for Future Research

Our experience creating the meta-analytic sample for this study revealed several issues in the intervention evaluation literature that limited our results and inspired our reflections regarding future directions for the field, detailed below.

Use Administrative Data to Examine Long-run Adult Effects of Existing RCTs

For the science of interventions to progress, we need more well-powered RCTs that measure adult follow-up effects (see Watts et al., 2019). Reflective of the state of the field, our study was limited by a small sample of studies with relatively small sample sizes. Our findings indicate that it is essential for studies to assess long-run adult follow-up regardless of the magnitude of posttest impacts or evidence of medium-term fadeout. A promising and cost-effective avenue for future research would be to examine the adult effects of existing RCTs that have not yet measured long-term effects using existing sources of administrative data. Assessing the long-run effects of programs is a promising direction for both applied educational researchers interested in policy and psychologists interested in basic questions about development.

Measure the Same Outcomes Across Studies and Assessment Waves

This literature was also hampered by inconsistent and biased measurement. 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. For example, in some cases, outcomes may have been over- or under-aligned to intervention content (Alvarez-Vargas et al., 2023; Halpin & Gilbert, 2024) and/or selectively measured or reported, especially at follow-up (Bailey et al., 2020). As a result, it is possible that intervention-level averages were biased in different directions that distorted important signal.

To address measurement issues, one fruitful direction for future research may be to establish a standardized set of outcomes that a large set of researchers consistently assess across studies and assessment waves, regardless of intervention targets. The battery should include measures that can be administered to children at different ages and reflect many of the core constructs interventions have attempted to change (e.g., achievement test scores), policy-relevant outcomes that may mediate intervention impacts (e.g., school grades and attendance), and individual difference measures perhaps not viewed as major intervention targets by most interventionists but which are thought to affect adult outcomes (e.g., conscientiousness). On the latter, for the field to advance, we believe that it is important to prioritize outcomes that carry strong construct validity and do not fall prey to the jingle-jangle fallacy (see Morrison & Grammer, 2017). Outcomes that can be assessed with administrative data can make follow-up cost-effective and protect against common method bias. Insofar as researchers have highly specific theories about other outcomes beyond those in the battery, they should additionally measure these outcomes. Critically, whatever measures researchers assess at one wave, they should prioritize assessing across follow-up waves.

Report All Effects in Consistent and Straightforward Ways

A third limitation of the current study is bias due to problematic reporting practices, including: a) reporting subgroup effects in lieu of main effects, b) using inconsistent methods to compute intervention impacts, and c) selectively reporting statistically significant results in the socially preferred direction. Inconsistent and selective reporting of intervention impacts makes it challenging to compare estimates across assessment waves. While we attempted to overcome problematic reporting practices through our coding approach, it is unlikely that our method overcame all biases.

Moving forward, we hope that researchers will report *main effects* on *all outcomes* that they assessed and use consistent pre-registered methods for computing impacts across outcomes and follow-up assessments. Insofar as moderator analyses are warranted, researchers should clearly specify their hypotheses *a priori* and be careful not to overweight subgroup estimates above main effects (see von Hippel & Schuetze, 2025). When changes in analytic methods are required across assessment waves, researchers should carefully document why the method changed and run their analyses using both updated and original methods, when possible.[10]

Match the Specificity of Analyses and Conclusions to the Specificity of Hypotheses

[10] For those interested in learning more about how to report intervention impacts, the What Works Clearinghouse provides useful guidelines.

Theoretical underdetermination is another issue that mixes unfavorably with inconsistent measurement and reporting to distort our understanding of longitudinal intervention impacts. Researchers' theories of change tend to be forgiving and difficult to falsify. Forgiving theory permits researchers to pick from a variety of theoretically compelling "stories" about an intervention's longitudinal impacts, flexibly fitting results to evolving theoretical explanations as a study carries on. Researcher stories can become increasingly distant from reality when anchored to potentially spurious, unlikely-to-replicate impacts (e.g., subgroup estimates, selective reporting). The lack of theoretical specificity implicit in the field may reflect both major issues in how psychological research is conducted (see Borsboom et al., 2021; Klein, 2014; Oberauer & Lewandowsky, 2019; Wagenmakers et al., 2012) and the real challenges of developing highly specific theories to explain human development (Eronen & Bringmann, 2021).

In future intervention evaluation work, we recommend that researchers match the specificity of their analyses and conclusions to the specificity of their hypotheses. First, we recommend that researchers pre-register their hypotheses and "forecasts" about how and why interventions will generate long-run effects. Researchers can use pre-registration to document their theories regardless of their specificity (i.e., if researchers do not have specific theories, they should document this). They can then pre-register analytic plans that are appropriately tailored to the specificity of their hypotheses. For example, if a research team does 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 and exploratory models that examine specific outcomes. For example, see Gelman and colleagues' (2009) practical approach to using Bayesian methods to estimate intervention impacts on many outcomes. After executing their pre-registered analyses, researchers should then match the theoretical and practical implications that they draw from the results to the specificity of their hypotheses to avoid over-interpreting the theoretical implications of results from spin-off analyses that may be unlikely to replicate.

When researchers run non-pre-registered models or examine outcomes that were not the focus of *a priori* theories, they should delineate the exploratory nature of such models (Wagenmakers et al., 2012). When spin-off analyses inspire new theory, researchers should make clear that their study was the basis for inspiration, rather than a falsification test. Researchers should be particularly cautious about how they discuss spin-off results when they paint a clear and compelling *post hoc* narrative.

Reform the Correlation-to-Intervention Pipeline

As detailed in Bailey et al. (2024), developmental theory suffers from an over-reliance on correlational evidence and an under-reliance on evidence from causally informative evaluations of interventions. Longitudinal intervention evaluations provide compelling tests of developmental theory by allowing researchers to causally manipulate important child skills and to observe how such changes affect long-run developmental trajectories. 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, 2020). Correlational studies lead to the expectation that variations in children's early skills are responsible for variations in individuals' long-run developmental trajectories. However, the current study replicates three other recent studies, finding that intervention-driven changes in children's skills fade considerably. Initial intervention impacts persist at a rate of approximately 40 to 50% at six months to one year after interventions end, and even less at one-to-two-year follow-up (Hart et al., 2024; Watts et al., 2024). Future research is needed to clarify the extent to which our estimates of the causal link between child skills and adult outcomes align with correlational estimates.

For researchers who are 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 attempts to probe the alignment between models fit to observational and (quasi-)experimental data. By eliminating sources of child-level endogeneity that are inherent in correlational models, examining links between interventions' initial and long-run intervention impacts is a compelling tool for assessing the effects of child functioning on later outcomes.[11] Beyond between-study examinations of the links between child skills and adult effects, which we relied on in the current study, within-study investigations of cluster-based RCTs are another promising approach to examining the long-run effects of exogenously generated variation in children's skills (Angrist et al., 2016; NCEERA, 2019).

Probe the Policy-Relevance of Adult Estimates

Depending on our modeling approach, we observed a meta-analytic average impact on adult outcomes ranging from .04 to .11 *SD.* Some researchers may be surprised that adult intervention impacts were as large as they were, whereas others may be shocked that impacts were not larger. Regarding the former, we did not find particularly concerning evidence for p-hacking or selection-into-adult-follow-up—perhaps reflective of our coding approach—and, thus, concluded that our adult impact estimates appeared to reflect real treatment-control group differences.

[11] It is worth noting that while models examining intervention impacts across time rely on exogenously generated variation, the models themselves are still correlational (i.e., they test the association between intervention impacts across time). As such, an association between intervention impacts on child skills and adult outcomes could reflect the influence of the specific child skill of focus or any number of correlated impacts on other factors (e.g., impacts on parents or other aspects of children's environments).

Regarding the latter, we anticipate that several factors may support the expectation of larger adult treatment effects. One challenge in forming a clear sense of the magnitude of adult impacts from the extant literature is that adult effects are often presented in the magnitude of the outcome of interest (e.g., percent increase in college completion). However, researchers may nonetheless intuit that the adult effects of programs are larger than .04 *SD*. And, for good reason: study authors often highlight impacts that are statistically differentiable from zero and, for an intervention with an original sample size of approximately 916 participants (the average observed in our sample) that observed 20% attrition (n = 733), an effect of about .21 *SD* would reach statistical significance. For a study with a smaller sample, or a study that had considerable attrition, statistical significance would only be detected for larger effects. By attempting to select for a relatively unbiased set of adult outcomes, and by averaging across all adult outcomes, our approach may have overcome reporting biases that have otherwise made it challenging to synthesize findings from the literature.

"Small" effects on important adult outcomes may still be meaningful and worth pursuing. As argued by Kraft (2020), a .04 *SD* effect can reflect a policy-relevant impact when benefits outweigh costs. Indeed, when benefit-cost ratios are greater than 1, small effect sizes can have non-trivial impacts on life course outcomes when considered at a population level. Considering intervention costs and scalability are essential to computing relative real-world benefits. Future work should evaluate the costs versus benefits and scalability of programs in our sample and determine how they stack up to alternative policy solutions that target life course outcomes.

Conclusion

The present study examined the long-run impacts of educational interventions. We identified 29 educational interventions that have evaluated the immediate and long-run effects of their programs through RCT designs. Across these 29 programs, we observed that interventions generated initial effects that faded rapidly in the medium-term. Programs generated long-run impacts on adult outcomes despite medium-term fadeout on child skills, but the trajectory of effects did not follow classic conceptualizations of “emergent” and “sleeper” effects. Instead, adult impacts were generally aligned with the magnitude of faded-out medium-term effects. Our analyses suggested that the mechanisms underlying these long-run effects are unclear. Indeed, we observed some, but usually not very strong, correspondence between initial intervention impacts on child skills and long-run impacts on adult outcomes, though our estimates were imprecise due to the small number of studies that have examined long-run effects. Together, it appears that there remains great opportunity for our field to continue to study how interventions targeting children come to shape life course trajectories. Current theories that predict strong skill-dependent paths connecting educational interventions to adult outcomes need revision. Future work could further test the Large Interconnected Network Theory, as our findings generally mapped onto its expectations. Additional examinations of the long-run adult effects of existing educational RCTs are needed to make further progress in this area. Continued investigations of the links between intervention-driven changes to child skills and later functioning can advance the practical implications of education research and refine basic developmental theory.

References

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

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

Angrist, J. D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., & Walters, C. R. (2016). Stand and deliver: Effects of Boston's charter high schools on college preparation, entry, and choice. *Journal of Labor Economics, 34*(2). https://doi.org/10.1086/683665

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

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

Bailey, D. H., Duncan, G. J., Watts, T., Clements, D. H., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *The 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

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

Bloom, B. S. (1964). *Stability and change in human characteristics.* John Wiley & Sons, Inc.

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

Borsboom, D., Van Der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756-766. https://doi.org/10.1177/174569162096964

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. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Harvard University Press.

Burchinal, M., & Vandell, D. L. (2025). School entry skills and young adult outcomes. *Early Childhood Research Quarterly*, *72*, 1–12. https://doi.org/10.1016/j.ecresq.2025.01.004

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

Capon, N., & Hulbert, J. (1973). The sleeper effect-an awakening. *Public Opinion Quarterly*, *37*(3), 333–358. https://doi.org/10.1086/268097

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. https://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

Coen, T., Nichols-Barrer, I., & Gleason, P. (2019). *Long-Term Impacts of KIPP Middle Schools on College Enrollment and Early College Persistence*. https://files.eric.ed.gov/fulltext/ED603636.pdf

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*.

Davis-Kean, P. E., Domina, T., Kuhfeld, M., Ellis, A., & Gershoff, E. T. (2022). It matters how you start: Early numeracy mastery predicts high school math course-taking and college attendance. *Infant and Child Development*, *31*(2), e2281. https://doi.org/10.1002/icd.2281

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

Dodge, K. A., Pettit, G. S., McClaskey, C. L., Brown, M. M., & Gottman, J. M. (1986). Social competence in children. *Monographs of the Society for Research in Child Development*, *51*(2), i–85. https://doi.org/10.2307/1165906

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

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

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., & Duckworth, K. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428-1446.

Duncan, G. J., Magnuson, K. A., & Ludwig, J. (2004). The endogeneity problem in developmental studies. *Research in Human Development*, 59–80.

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

Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences, 114*(14), 3714-3719. https://doi.org/ 10.1073/pnas.1618569114

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

Gelman, A., Hill, J., & Yajima, M. (2009). *Why we (usually) don't have to worry about multiple comparisons* (No. arXiv:0907.2478). arXiv. https://doi.org/10.48550/arXiv.0907.2478

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

Halpin, P., & Gilbert, J. (2024). *Testing whether reported treatment effects are unduly dependent on the specific outcome measure used* (No. arXiv:2409.03502). arXiv. https://doi.org/10.48550/arXiv.2409.03502

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

Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research*, *66*(2), 99–136. https://doi.org/10.3102/00346543066002099

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, *19*(4), 451–464. https://doi.org/10.1016/j.labeco.2012.05.014

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.

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

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. (Studies in social psychology in World War II), Vol. 3* (pp. x, 345). Princeton University Press.

Hutmacher, F., & Franz, D. J. (2025). Approaching psychology’s current crises by exploring the vagueness of psychological concepts: Recommendations for advancing the discipline. *American Psychologist*, *80*(2), 220–231. https://doi.org/10.1037/amp0001300

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. https://doi.org/10.1023/A:1022137920532

Ioannidis, J. P., Stanley, T. D., and Doucouliagos, H. (2017). The power of bias in economics research. *Economic Journal, 127*(605):F236–F265. https://doi.org/10.1111/ecoj.12461

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

Kagan, J., & Moss, H. (1962). *Birth to maturity: A study in psychological development*. New York: John Wiley and Sons Inc. http://archive.org/details/birthtomaturitys0000kaga_s4x0

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? Theory & Psychology, 24(3), 326-338. https://doi.org/10.1177/0959354314529616

Koepp, A. E., Watts, T. W., Gershoff, E. T., Ahmed, S. F., Davis-Kean, P., Duncan, G. J., Kuhfeld, M., & Vandell, D. L. (2023). Attention and behavior problems in childhood predict adult financial status, health, and criminal activity: A conceptual replication and extension of Moffitt et al. (2011) using cohorts from the United States and the United Kingdom. *Developmental Psychology*, *59*(8), 1389–1406. https://doi.org/10.1037/dev0001533

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

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

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science, 11*(5), 730–749. https://doi.org/10.1177/1745691616662243

Mischel, W. (2008). The toothbrush problem. *APS observer*, 21(11), 1-3. https://www.psychologicalscience.org/observer/the-toothbrush-problem

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, W. M., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, *108*(7), 2693–2698. https://doi.org/10.1073/pnas.1010076108

Morey, R. D., Davis-Stober, C. P. (2025). On the poor statisitical properties of the p-curve meta-analytic procedure. *Journal of the American Statistical Association*.

Morrison, F. J., & Grammer, J. K. (2016). Conceptual clutter and measurement mayhem: Proposals for cross-disciplinary integration in conceptualizing and measuring executive function. In *Executive function in preschool-age children: Integrating measurement, neurodevelopment, and translational research* (pp. 327–348). American Psychological Association. https://doi.org/10.1037/14797-015

NCEERA. (2019). *Do charter middle schools improve students' college outcomes?* https://ies.ed.gov/ncee/2025/01/20194005highlights-pdf

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review, 26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

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

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

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*(Suppl 1), 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

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor. In D. B. Grusky & J. Hill (Eds.), *Inequality in the 21st Century* (1st ed., pp. 177–189). Routledge. https://doi.org/10.4324/9780429499821-33

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science*, *24*(7), 1301–1308. https://doi.org/10.1177/0956797612466268

Rohrer, J., & Lucas, R. (2020). *Causal effects of well-being on health: It's complicated*. OSF. https://doi.org/10.31234/osf.io/wgbe4

Rosengarten, M. L., Hart, E. R., Bailey, D. H., McCormick, M. P., Lovett, B. J., & Watts, T. W. (2024). *Using meta-analytic data to examine fadeout and persistence of intervention impacts on constrained and unconstrained skills. EdWorkingPaper No. 24-1069*. Annenberg Institute for School Reform at Brown University. https://eric.ed.gov/?id=ED672296

Sarigiani, P. A., & Spierling, T. (2011). Sleeper effect of divorce. In *Encyclopedia of Child Behavior and Development* (pp. 1378–1385). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-79061-9_2666

Schneider-Rosen, K., & Cicchetti, D. (1984). The relationship between affect and cognition in maltreated infants: Quality of attachment and the development of visual self-recognition. *Child Development*, *55*(2), 648–658. https://doi.org/10.2307/1129976

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. https://doi.org/10.1177/105381518100400105

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., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, *62*(4), 273–296. https://doi.org/10.1016/j.cogpsych.2011.03.001

Simonsohn, U. (2025, September 23). *P-curve works in practice, but would it work if you dropped a piano on it?* Data Colada. https://datacolada.org/129

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

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

Thompson, M. J., Hinnant, J. B., Erath, S. A., & El-Sheikh, M. (2024). The legacy of harsh parenting: Enduring and sleeper effects on trajectories of externalizing and internalizing symptoms. *Developmental Psychology*, *60*(8), 1482–1499. https://doi.org/10.1037/dev0001754

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

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

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

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

Wan, S., Bond, T. N., Lang, K., Clements, D. H., Sarama, J., & Bailey, D. H. (2021). Is intervention fadeout a scaling artefact? *Economics of Education Review*, *82*, 102090.

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

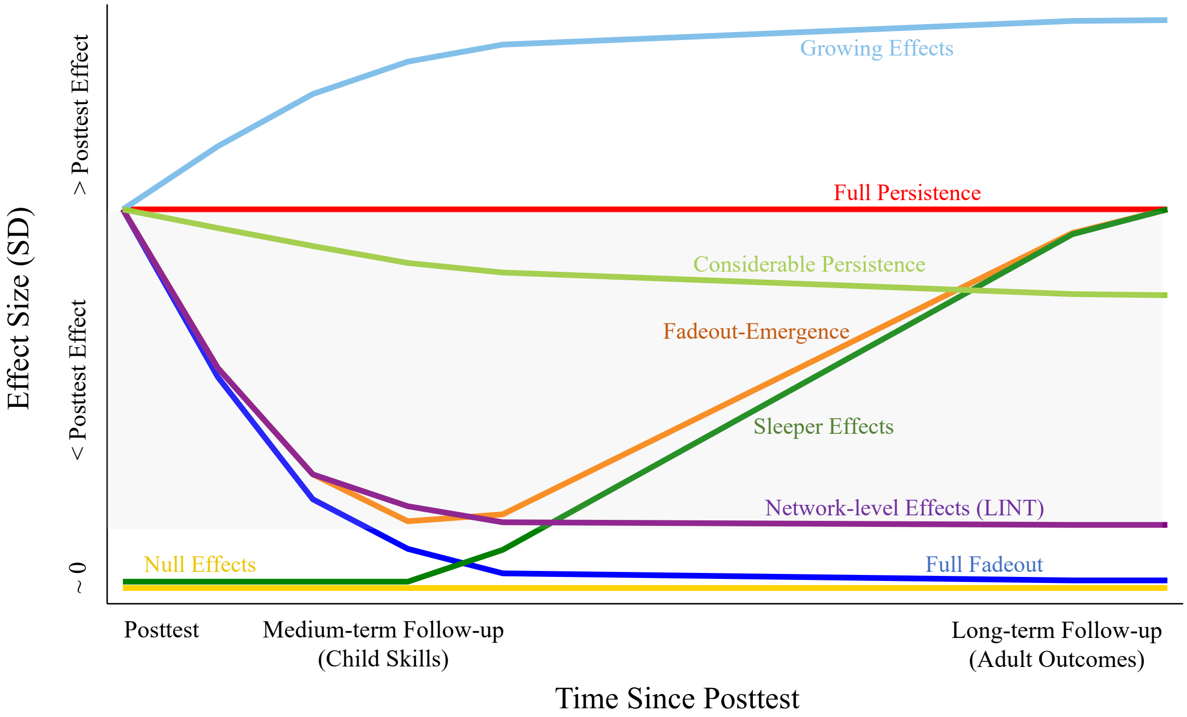
Watts, T. W., Botvin, C. M., Bailey, D. H., Hart, E. R., Mattera, S., Clements, D. H., Sarama, J., Farran, D., & Lipsey, M. W. (2024). *Predicting persistence and fadeout across multisite RCTs of an early childhood mathematics curriculum intervention* [Manuscript under review]. University of Chicago Voices. https://bpb-us-w2.wpmucdn.com/voices.uchicago.edu/dist/a/1176/files/2025/04/Watts-et-al.-under-review-BB-Multisite-Fadeout.pdf

Woodhead, M. (1988). When psychology informs public policy. *American Psychologist*.

Zigler, E., & Berman, W. (1983). Discerning the future of early childhood intervention. *American Psychologist*, *38*(8), 894–906.

Figure 1

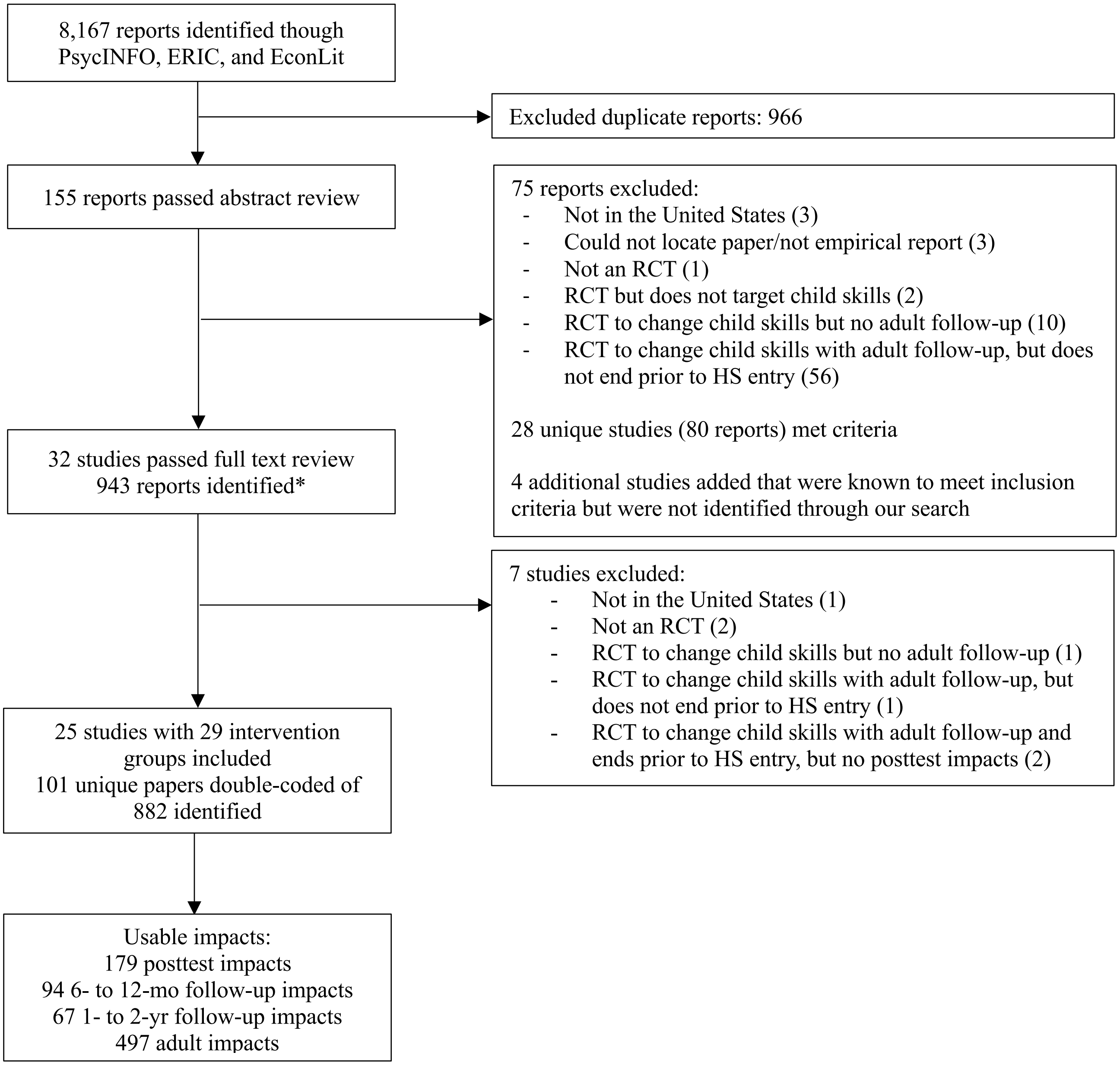
Theoretical Trajectories of Intervention Impacts



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 Reports and Studies into the Meta-Analytic Sample


8,167 reports identified though PsycINFO, ERIC, and EconLit
Excluded duplicate reports: 966
155 reports passed abstract review
75 reports excluded:
- Not in the United States (3)
- Could not locate paper/not empirical report (3)
- Not an RCT (1)
- RCT but does not target child skills (2)
- RCT to change child skills but no adult follow-up (10)
- RCT to change child skills with adult follow-up, but does not end prior to HS entry (56)
28 unique studies (80 reports) met criteria
4 additional studies added that were known to meet inclusion criteria but were not identified through our search
32 studies passed full text review 943 reports identified*
7 studies excluded:
- Not in the United States (1)
- Not an RCT (2)
- RCT to change child skills but no adult follow-up (1)
- RCT to change child skills with adult follow-up, but does not end prior to HS entry (1)
- RCT to change child skills with adult follow-up and ends prior to HS entry, but no posttest impacts (2)
25 studies with 29 intervention groups included 101 unique papers double-coded of 882 identified
Usable impacts:
179 posttest impacts
94 6- to 12-mo follow-up impacts
67 1- to 2-yr follow-up impacts
497 adult impacts

Notes: *After identifying that one of the studies was conducted outside of the United States early in the search process, we stopped searching for additional reports for this study. ** We identified 882 reports by summing the number of unique reports that we identified for each intervention and 101 by summing the number of reports we coded for each intervention. Notably, this figure does not account for the fact that the same report may have been identified (or coded) for multiple interventions.

Table 1
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

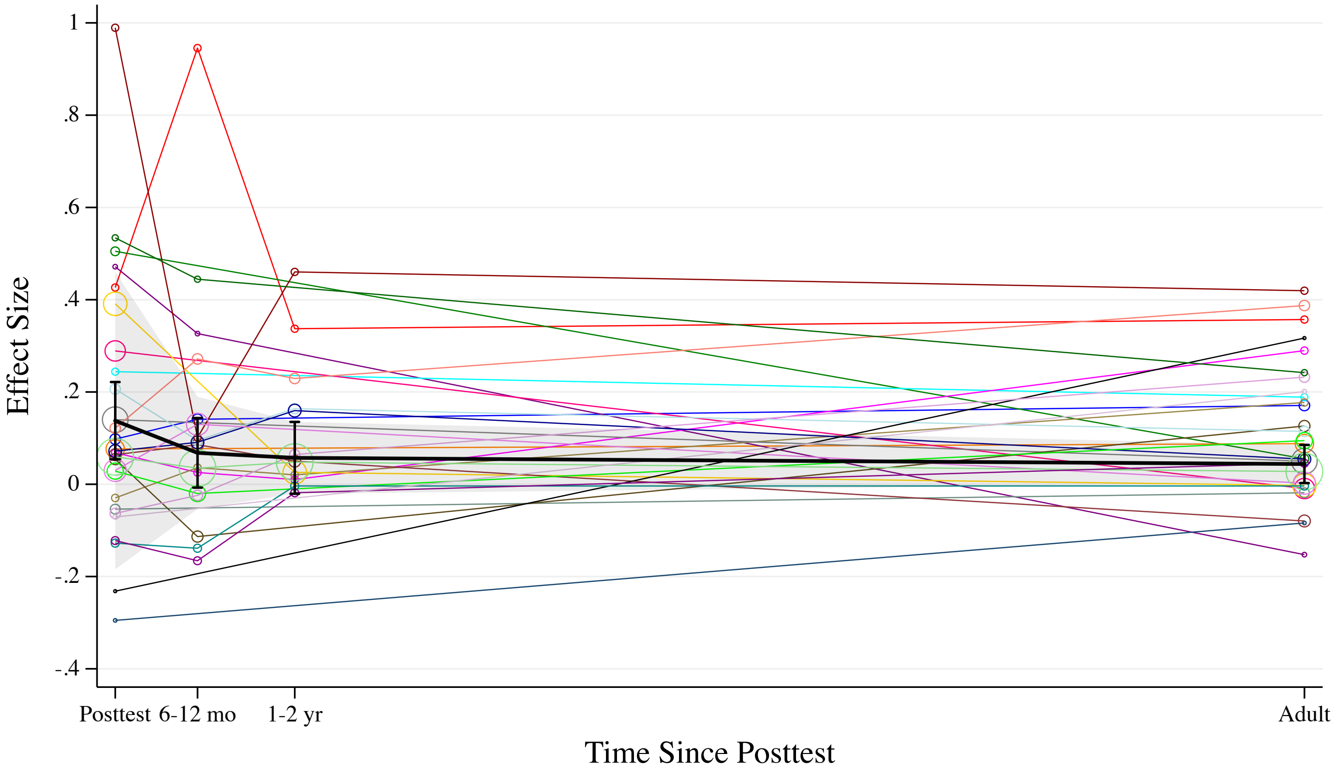
Notes: This table provides information on the outcomes that compose the posttest and adult impacts in our sample. All meta-analytic averages were estimated using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level).
*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

Table 2
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

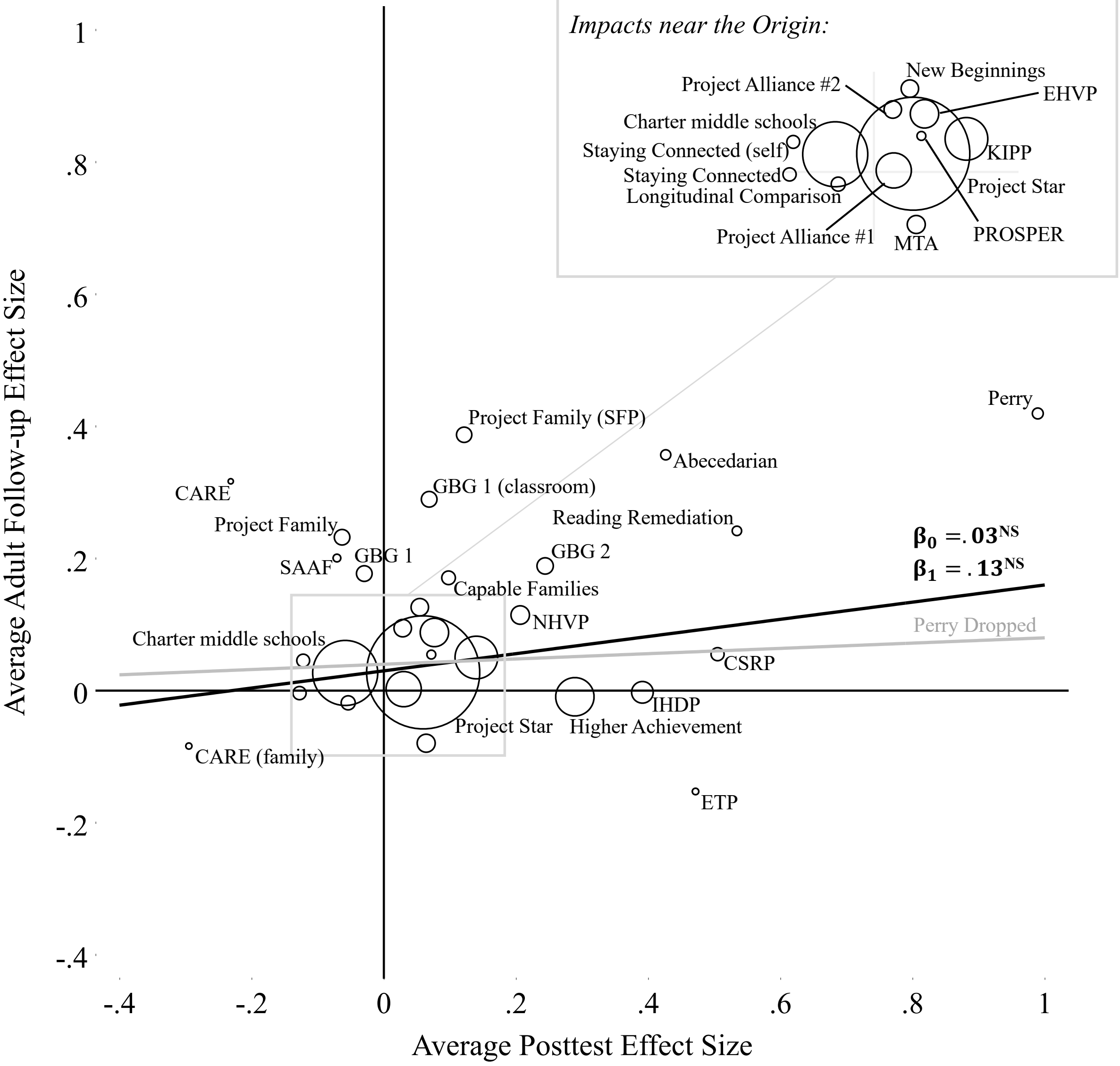
Notes: All meta-analytic averages were estimated using inverse sampling variance weighting, a random effect for study, and cluster-robust standard errors (with clustering at the study level). Negative I^2 statistics were rounded to zero. "Aligned groups" were cases where the same intervention assessed 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 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 2, 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 S1 for full intervention names associated with the label abbreviations.
* $p < .05$, ** $p < .01$, *** $p < .001$

Supplemental Information for:
Childhood Interventions and Life Course Development

Additional Methodological Details

Protocol

The protocol that guided the study team's completion of various steps in the inclusion-exclusion process and coding will be made available on LDbase prior to publication (https://doi.org/10.33009/ldbase.1719529626.152e). The protocol largely followed the logic of that used in creating the original Meta-analysis of Educational RCTs with Follow-up (MERF) sample (see Hart et al., 2024).

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 the team 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

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 reports 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.

Backwards/Forwards Search Process

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

Effect Size and Standard Error Computations

Figure S1 details how effect sizes were computed based on the reported information.

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 Sensitivity Checks

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.

Age and Timing Issues

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 S11).

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 S12).

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 S12) that dropped studies with pre-end-of-treatment posttests.

“Duplicate” Issue

When coding the posttest and short-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 S12).

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 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 S12) 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 S12) 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.

Table S1
Descriptives of the Included Interventions

Study ~ Treatment	Start year	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	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	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	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 et al., 1986, 1994)	1979	193	28	prenatal	x	x	x	Reduce child developmental problems in families with young, unmarried, and/or low-SES mothers through nurse home visiting targeting maternal wellbeing, parenting (e.g., maltreatment), and economic self-sufficiency.	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	Avg. *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies)*
Project Carolina Approach to Responsive Education (CARE) ~ Center-based plus Family Education Group (Campbell et al., 2008, 2013)	1979	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 status.
Project Carolina Approach to Responsive Education (CARE) ~ Family Education Group (Campbell et al., 2008, 2013)	1979	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) (Cole et al., 1993; Jenkins et al., 2006)	1983	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 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

Study ~ Treatment	Start year	Avg. *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies*)
									(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	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	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 social-emotional outcomes measured at posttest in addition to test scores.
Good Behavior Game 2 (GBG 2)	1986	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

Study ~ Treatment	Start year	Avg. *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies*)
(Dolan et al., 1993; Kellan, 2008; Poduska et al., 2008)									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., 2021; Kitzman et al., 1997, 2019; Olds et al., 2014)	1991	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.
Multimodal Treatment Study of ADHD (MTA) (Hechtman et al., 2005; Jensen et al.,	1992	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.

Study ~ Treatment	Start year	Avg. *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies)*
2007; Langberg et al., 2010; MTA, 1999, 2004; Swanson et al., 2017; Wells et al., 2006)									
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	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.
Good Behavior Game (GBG)~ Classroom (Bradshaw et al., 2009; Ialongo et al., 1999; Musci et al., 2018; Wang et al., 2012;	1993	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.

Study ~ Treatment	Start year	Avg. *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies)*
Saunders, 2007)									
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	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 al., 2006, 2008a, 2009a)	1993	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.
Project Family ~ Strengthening Families Group (Spoth et al., 1999, 2002, 2006, 2008a, 2008b, 2009a,	1993	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.

Study ~ Treatment	Start year	Avg. *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies*)
2009b, 2013; Trudeau et al., 2012)									
Project Alliance #1 (Gardner et al., 2008; Panza, 2019; Van Ryzin et al., 2013; Van Ryzin & Dishion, 2012; Zhang et al., 2024)	1997	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 intervention 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	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 Families Program (Spoth et al., 2002, 2005, 2006, 2008, 2013; Trudeau et al., 2003, 2016)	1999	1372	16	148	x	x		Prevent early substance use initiation through universal program 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 treated the end of seventh grade as the end of the intervention. Additionally, while not reported consistently across papers, the authors did note that there were 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.

Study ~ Treatment	Start year	Avg. *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies)*
PROSPER Family- and School-focused Intervention (Redmond et al., 2009; Spoth et al., 2011, 2013, 2017, 2022)	2003	10170	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	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.
Charter Middle Schools (Clark et al., 2015; NCEE, 2019)	2006	2030	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	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.

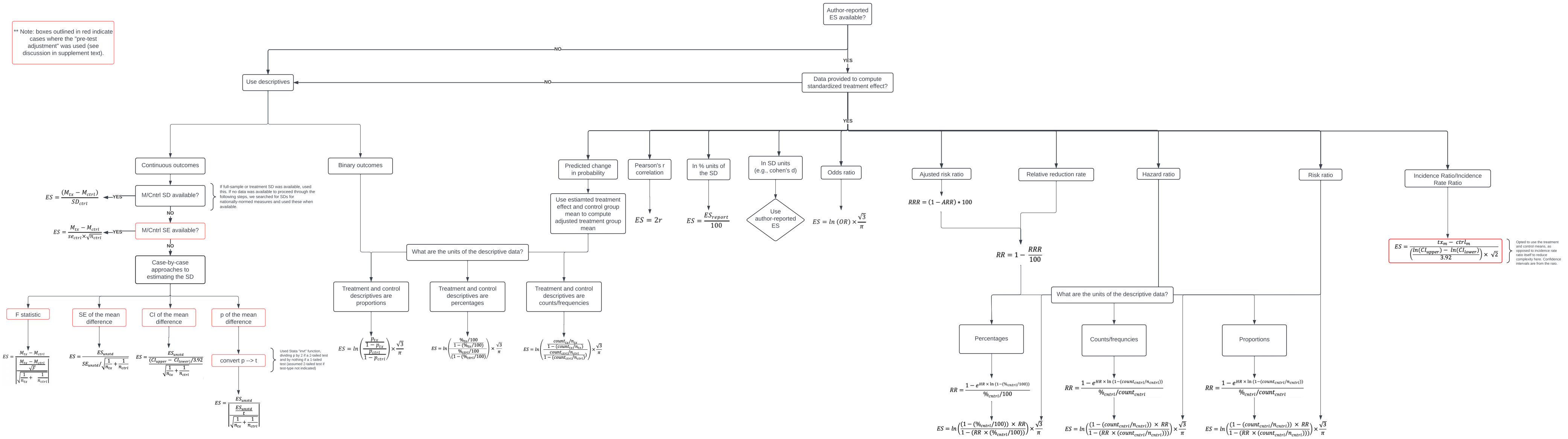
Study ~ Treatment	Start year	Avg. *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies)*
Knowledge is Power Program (KIPP) (Coen et al., 2019; Demers et al., 2023; Tuttle et al., 2013)	2009	590	38	126	x	x	x	Improve 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	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.
Staying Connected ~ Parent and Adolescent Administered (Haggerty et al., 2007, 2015)	Early 2000s *	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	Early 2000s *	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.

Study ~ Treatment	Start year	Avg. *n*	Duration (mos)	Age (mos)	Parents	Soc-emo	Cog	Standout intervention features	Reporting distinctions (*Notable differences or idiosyncrasies)*
(Haggerty et al., 2007, 2015)									
Strong African American Families (SAAF) (Brody et al., 2020; Miller et al., 2014)	Late 90s-Early 2000s *	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**	**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.

Figure S1

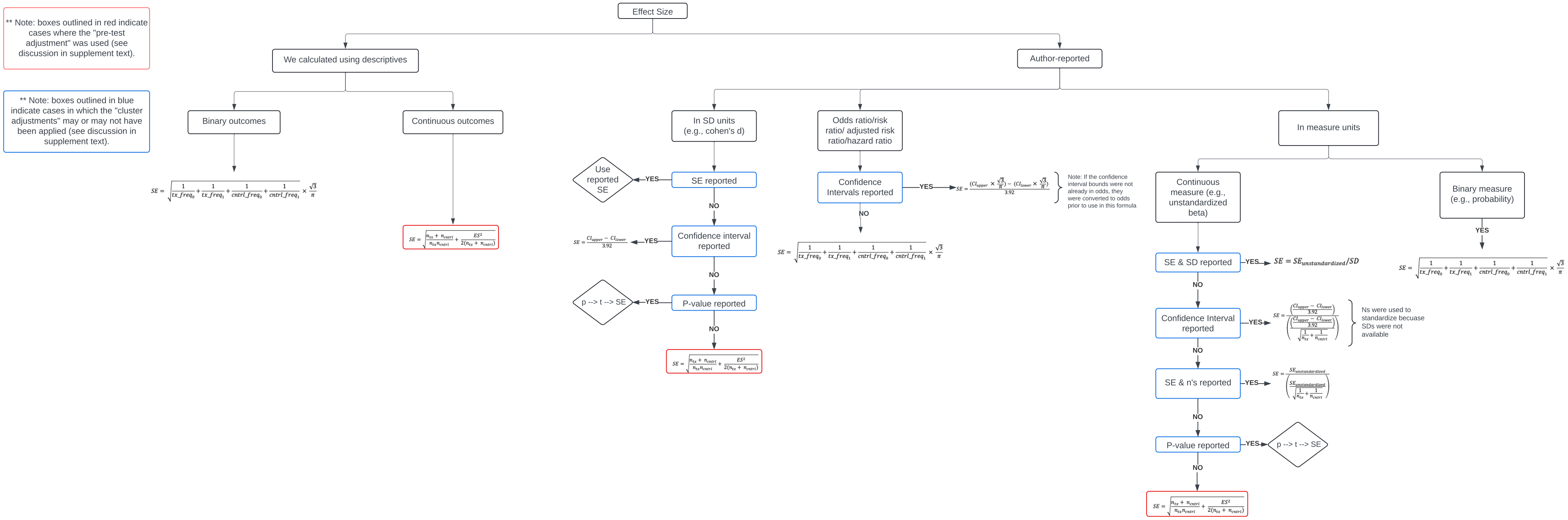
Effect Size Calculation Decision Flow


** Note: boxes outlined in red indicate cases where the "pre-test adjustment" was used (see discussion in supplement text).
Author-reported ES available?
NO
YES
Data provided to compute standardized treatment effect?
NO
YES
Use descriptives
Continuous outcomes
Binary outcomes
M/Cntrl SD available?
YES
$ES = \frac{(M_{tx} - M_{ctrl})}{SD_{ctrl}}$
If full-sample or treatment SD was available, used this. If no data was available to proceed through the following steps, we searched for SDs for nationally-normed measures and used these when available.
NO
M/Cntrl SE available?
YES
$ES = \frac{M_{tx} - M_{ctrl}}{se_{ctrl} \times \sqrt{n_{ctrl}}}$
NO
Case-by-case approaches to estimating the SD
F statistic
SE of the mean difference
CI of the mean difference
p of the mean difference
$ES = \frac{M_{tx} - M_{ctrl}}{\frac{\frac{M_{tx} - M_{ctrl}}{\sqrt{F}}}{\sqrt{\frac{1}{n_{tx}} + \frac{1}{n_{ctrl}}}}}$
$ES = \frac{ES_{unstd}}{SE_{unstd} / \sqrt{\frac{1}{n_{tx}} + \frac{1}{n_{ctrl}}}}$
$ES = \frac{ES_{unstd}}{\frac{(CI_{upper} - CI_{lower})/3.92}{\sqrt{\frac{1}{n_{tx}} + \frac{1}{n_{ctrl}}}}}$
convert p --> t
Used Stata "invt" function, dividing p by 2 if a 2-tailed test and by nothing if a 1-tailed test (assumed 2-tailed test if test-type not indicated)
$ES = \frac{ES_{unstd}}{\frac{\frac{ES_{unstd}}{t}}{\sqrt{\frac{1}{n_{tx}} + \frac{1}{n_{ctrl}}}}}$
Predicted change in probability
Use estiamted treatment effect and control group mean to compute adjusted treatment group mean
What are the units of the descriptive data?
Treatment and control descriptives are proportions
Treatment and control descriptives are percentages
Treatment and control descriptives are counts/frequencies
$ES = ln\left(\frac{\frac{p_{tx}}{1 - p_{tx}}}{\frac{p_{ctrl}}{1 - p_{ctrl}}}\right) \times \frac{\sqrt{3}}{\pi}$
$ES = ln\left(\frac{\frac{\%_{tx}/100}{1 - (\%_{tx}/100)}}{\frac{\%_{ctrl}/100}{(1 - (\%_{ctrl}/100)}}\right) \times \frac{\sqrt{3}}{\pi}$
$ES = ln\left(\frac{\frac{count_{tx}/n_{tx}}{1 - (count_{tx}/n_{tx})}}{\frac{count_{ctrl}/n_{ctrl}}{1 - (count_{ctrl}/n_{ctrl})}}\right) \times \frac{\sqrt{3}}{\pi}$
Pearson's r correlation
$ES = 2r$
In % units of the SD
$ES = \frac{ES_{report}}{100}$
In SD units (e.g., cohen's d)
Use author-reported ES
Odds ratio
$ES = ln\ (OR) \times \frac{\sqrt{3}}{\pi}$
Ajusted risk ratio
$RRR = (1 - ARR) * 100$
Relative reduction rate
$RR = 1 - \frac{RRR}{100}$
Hazard ratio
Risk ratio
Incidence Ratio/Incidence Rate Ratio
$ES = \frac{tx_m - ctrl_m}{\left(\frac{ln(CI_{upper}) - ln(CI_{lower})}{3.92}\right) \times \sqrt{2}}$
Opted to use the treatment and control means, as opposed to incidence rate ratio itself to reduce complexity here. Confidence intervals are from the raio.
What are the units of the descriptive data?
Percentages
Counts/frequncies
Proportions
$RR = \frac{1 - e^{HR \times \ln(1 - (\%_{cntrl}/100))}}{\%_{cntrl}/100}$
$RR = \frac{1 - e^{HR \times \ln(1 - (count_{cntrl}/n_{cntrl}))}}{\%_{cntrl}/count_{cntrl}}$
$RR = \frac{1 - e^{HR \times \ln(1 - (count_{cntrl}/n_{cntrl}))}}{\%_{cntrl}/count_{cntrl}}$
$ES = ln\left(\frac{(1 - (\%_{cntrl}/100)) \times RR}{1 - (RR \times (\%_{cntrl}/100))}\right) \times \frac{\sqrt{3}}{\pi}$
$ES = ln\left(\frac{(1 - (count_{cntrl}/n_{cntrl})) \times RR}{1 - (RR \times (count_{cntrl}/n_{cntrl}))}\right) \times \frac{\sqrt{3}}{\pi}$
$ES = ln\left(\frac{(1 - (count_{cntrl}/n_{cntrl})) \times RR}{1 - (RR \times (count_{cntrl}/n_{cntrl}))}\right) \times \frac{\sqrt{3}}{\pi}$

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

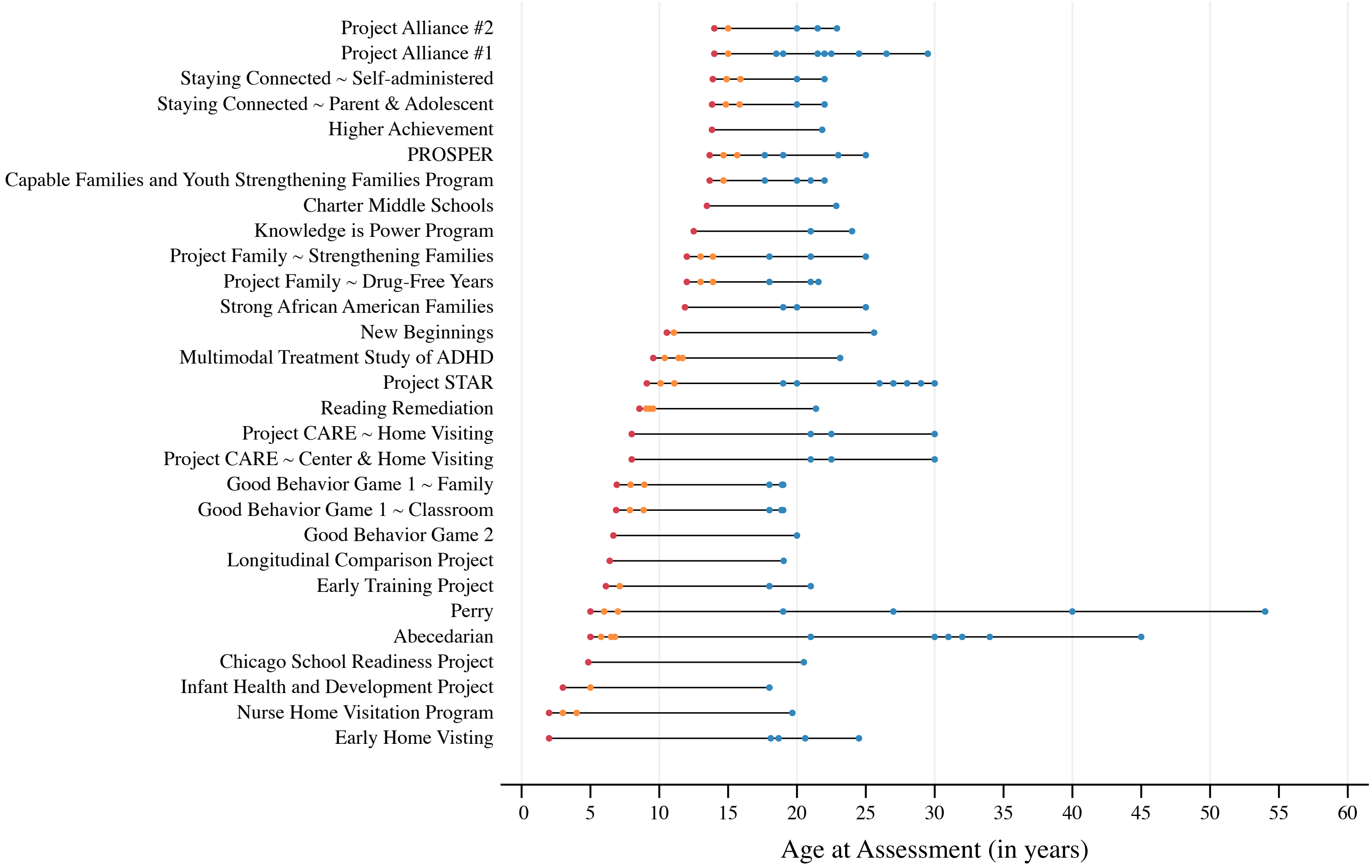
Alternative Standard Errors Calculation Decision Flow


** Note: boxes outlined in red indicate cases where the "pre-test adjustment" was used (see discussion in supplement text).
** Note: boxes outlined in blue indicate cases in which the "cluster adjustments" may or may not have been applied (see discussion in supplement text).
Effect Size
We calculated using descriptives
Author-reported
Binary outcomes
Continuous outcomes
$SE = \sqrt{\frac{1}{tx\_freq_0} + \frac{1}{tx\_freq_1} + \frac{1}{cntrl\_freq_0} + \frac{1}{cntrl\_freq_1}} \times \frac{\sqrt{3}}{\pi}$
$SE = \sqrt{\frac{n_{tx} + n_{cntrl}}{n_{tx}n_{cntrl}} + \frac{ES^2}{2(n_{tx} + n_{cntrl})}}$
In SD units (e.g., cohen's d)
SE reported
YES
Use reported SE
NO
Confidence interval reported
YES
$SE = \frac{CI_{upper} - CI_{lower}}{3.92}$
NO
P-value reported
YES
p --> t --> SE
NO
$SE = \sqrt{\frac{n_{tx} + n_{cntrl}}{n_{tx}n_{cntrl}} + \frac{ES^2}{2(n_{tx} + n_{cntrl})}}$
Odds ratio/risk ratio/ adjusted risk ratio/hazard ratio
Confidence Intervals reported
YES
$SE = \frac{(CI_{upper} \times \frac{\sqrt{3}}{\pi}) - (CI_{lower} \times \frac{\sqrt{3}}{\pi})}{3.92}$
Note: If the confidence interval bounds were not already in odds, they were converted to odds prior to use in this formula
NO
$SE = \sqrt{\frac{1}{tx\_freq_0} + \frac{1}{tx\_freq_1} + \frac{1}{cntrl\_freq_0} + \frac{1}{cntrl\_freq_1}} \times \frac{\sqrt{3}}{\pi}$
In measure units
Continuous measure (e.g., unstandardized beta)
Binary measure (e.g., probability)
YES
$SE = \sqrt{\frac{1}{tx\_freq_0} + \frac{1}{tx\_freq_1} + \frac{1}{cntrl\_freq_0} + \frac{1}{cntrl\_freq_1}} \times \frac{\sqrt{3}}{\pi}$
SE & SD reported
YES
$SE = SE_{unstandardized}/SD$
NO
Confidence Interval reported
YES
$SE = \frac{\left(\frac{CI_{upper} - CI_{lower}}{3.92}\right)}{\left(\frac{\left(\frac{CI_{upper} - CI_{lower}}{3.92}\right)}{\sqrt{\frac{1}{n_{tx}} + \frac{1}{n_{cntrl}}}}\right)}$
Ns were used to standardize becuase SDs were not available
NO
SE & n's reported
YES
$SE = \frac{SE_{unstandardized}}{\left(\frac{SE_{unstandardized}}{\sqrt{\frac{1}{n_{tx}} + \frac{1}{n_{cntrl}}}}\right)}$
NO
P-value reported
YES
p --> t --> SE
NO
$SE = \sqrt{\frac{n_{tx} + n_{cntrl}}{n_{tx}n_{cntrl}} + \frac{ES^2}{2(n_{tx} + n_{cntrl})}}$

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

Figure S3

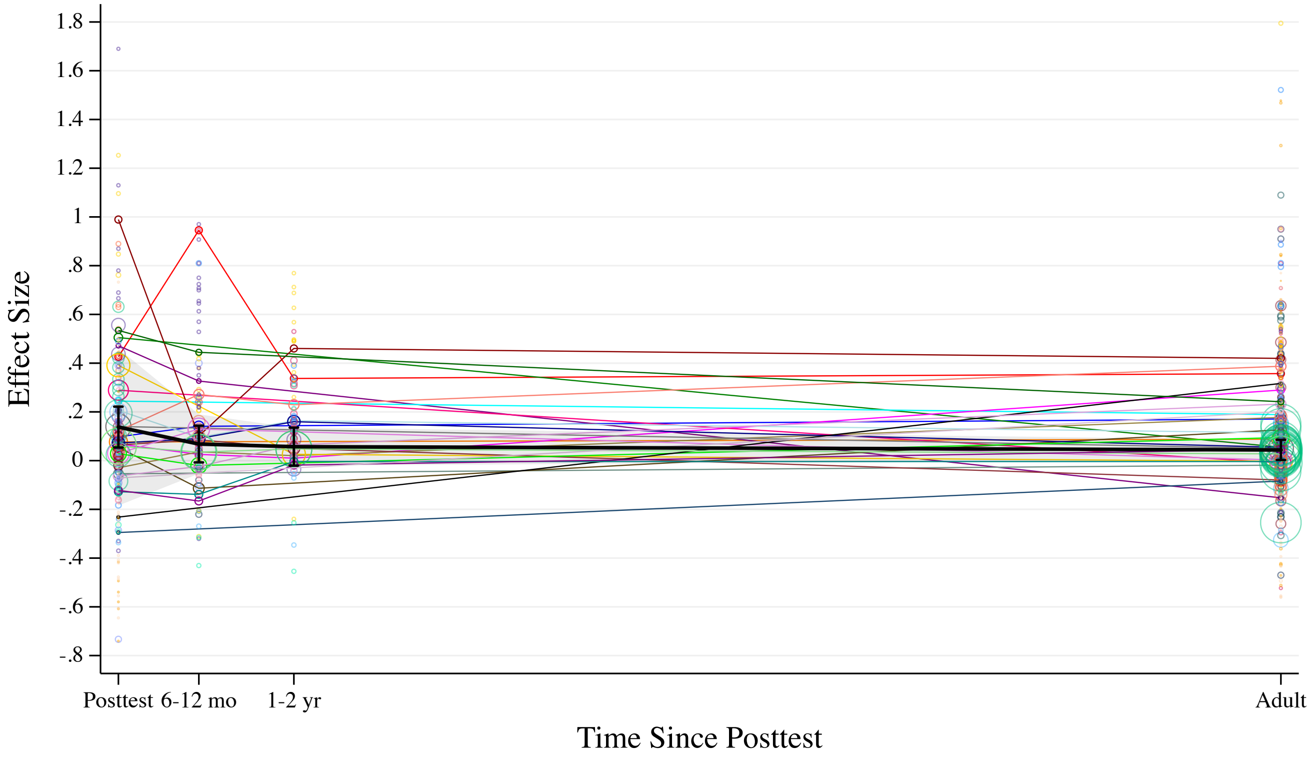
Intervention-Level Assessment Timelines



Notes: Each line portrays the assessments collected for each intervention in the sample. Red coordinates reflect posttest assessments. Orange coordinates reflect short-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 S2
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 among posttest and short-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 nested within a random effect for study, and cluster-robust standard errors (with clustering at the study level).
* $p < .05$, ** $p < .01$, *** $p < .001$

Table S3
Adult Follow-Up Impacts Predicted by Posttest and Short-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 short-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 short-term follow-up impacts and adult impacts using only "aligned groups" in which the outcomes and instruments were consistent across posttest and short-term follow-up waves. Negative I^2 statistics were rounded to zero.

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

Table S4

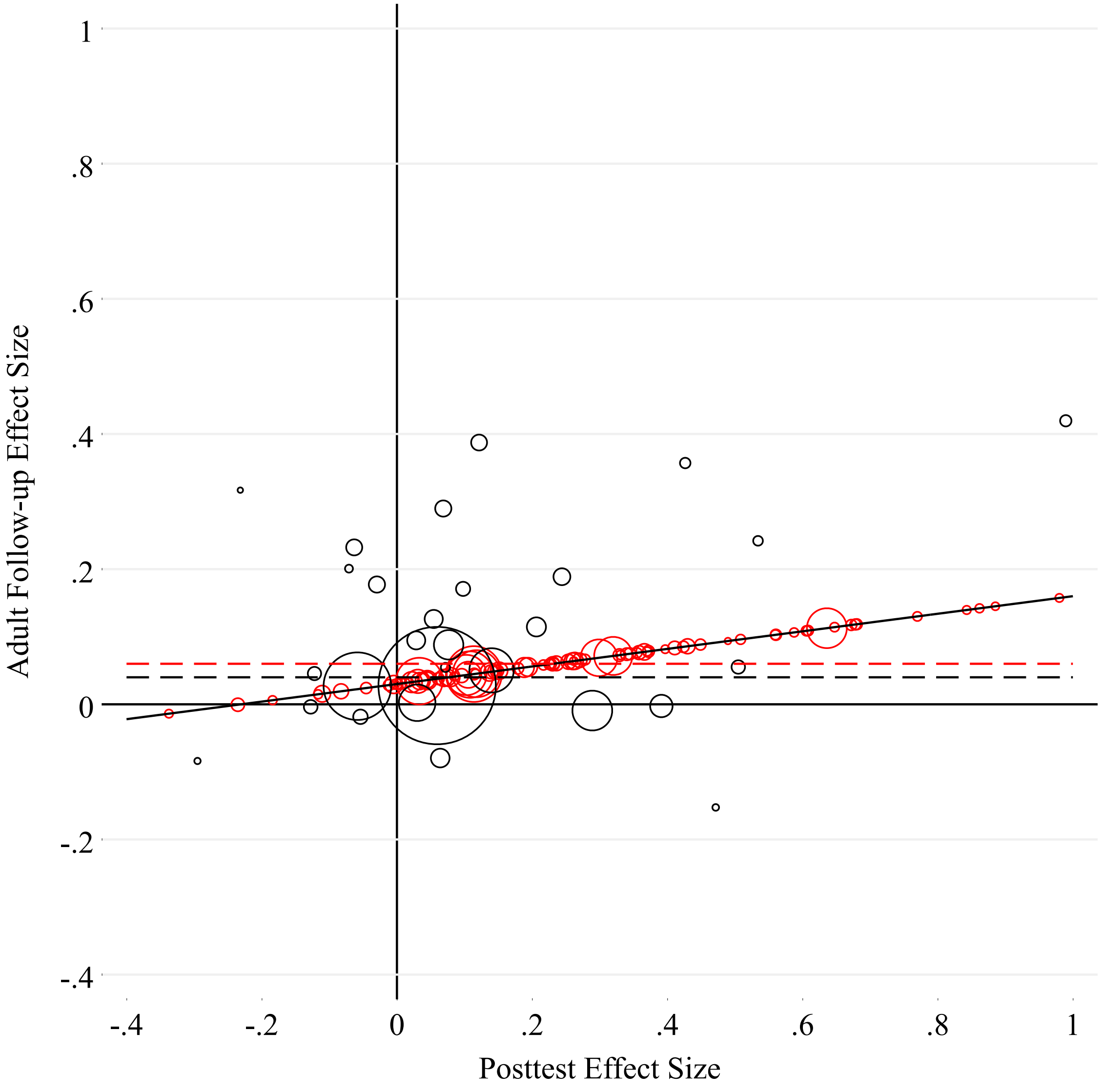
Adult Follow-Up Impacts Predicted by Posttest Impacts with Each Study Dropped

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).

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

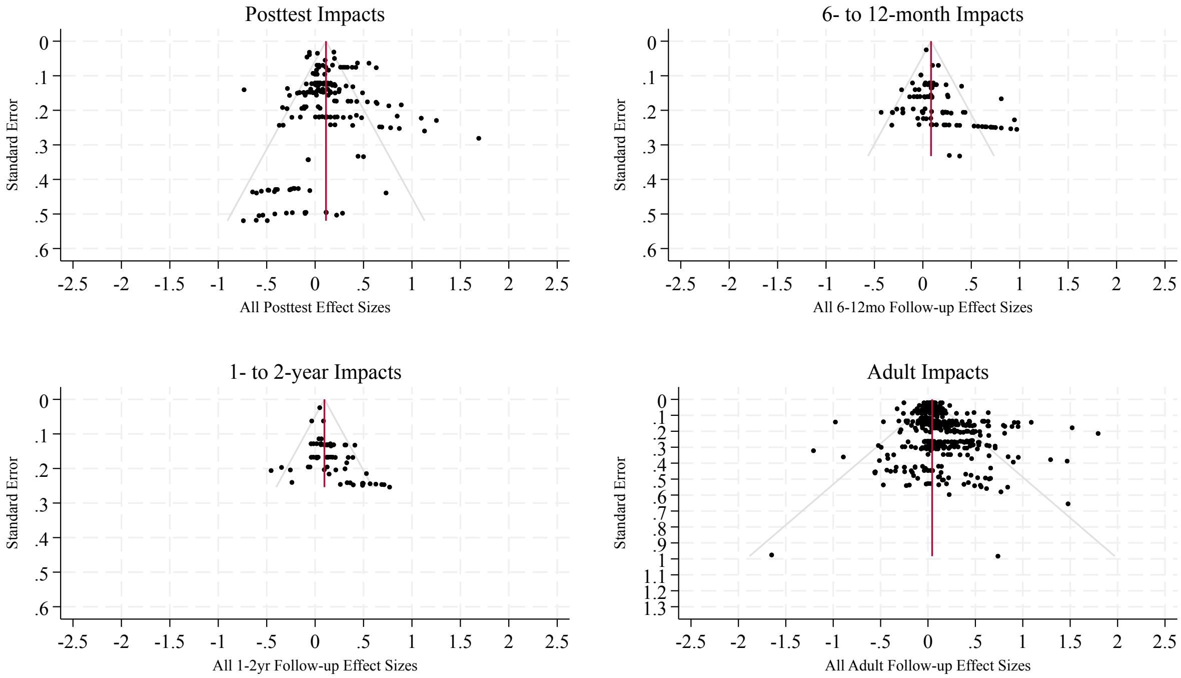
Figure S5
Using the MERF-E Estimates to Forecast the Long-Run Effects of Other Interventions



Notes: The black coordinates represent the average posttest and adult follow-up impact for each intervention in the MERF-E sample (n = 29 intervention groups). The black solid line reflects the association between average posttest and adult follow-up effects in the MERF-E 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 original 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 MERF-E 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 the MERF-E sample, the adult impact variances are used. For imputed adult impacts from the original 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 original MERF sample that had average posttest impacts that were greater than 1.

Figure S6

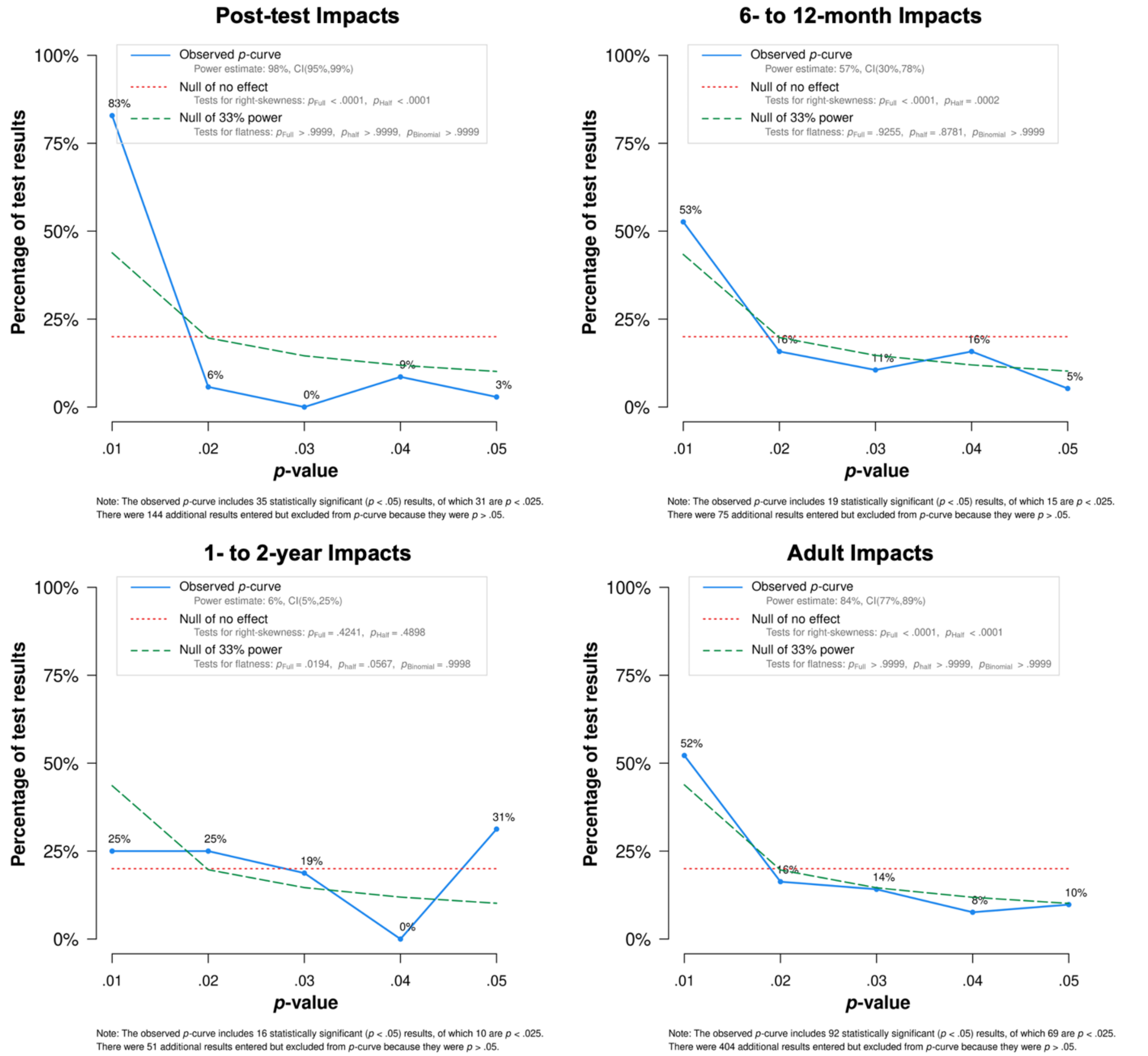
Funnel Plots for Posttest, Short-Term Follow-Up, and Adult Follow-Up Impacts



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 short-term follow-up impacts.

Figure S7

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



Note: The observed *p*-curve includes 35 statistically significant ($p < .05$) results, of which 31 are $p < .025$. There were 144 additional results entered but excluded from *p*-curve because they were $p > .05$.

Note: The observed *p*-curve includes 19 statistically significant ($p < .05$) results, of which 15 are $p < .025$. There were 75 additional results entered but excluded from *p*-curve because they were $p > .05$.

Note: The observed *p*-curve includes 16 statistically significant ($p < .05$) results, of which 10 are $p < .025$. There were 51 additional results entered but excluded from *p*-curve because they were $p > .05$.

Note: The observed *p*-curve includes 92 statistically significant ($p < .05$) results, of which 69 are $p < .025$. There were 404 additional results entered but excluded from *p*-curve because they were $p > .05$.

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 S5

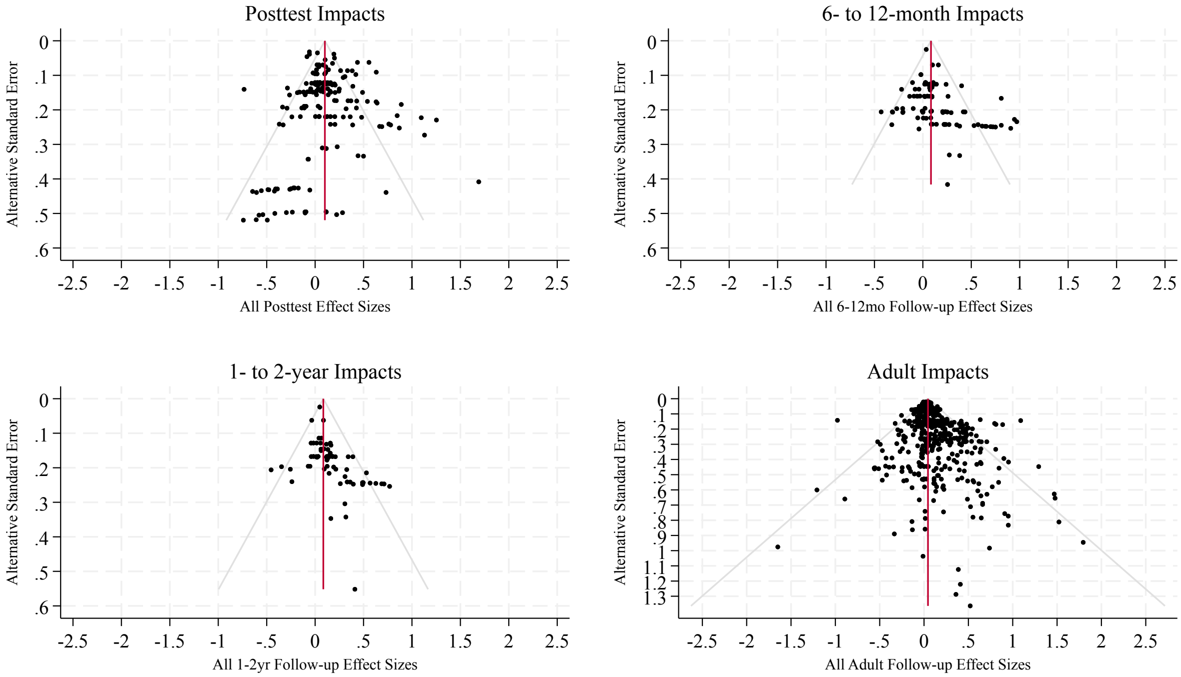
Primary Results Using Alternative Weighting Approaches and the Correlated-and-Hierarchical Model

	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, short-term follow-up, and adult impacts and the association among 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.

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

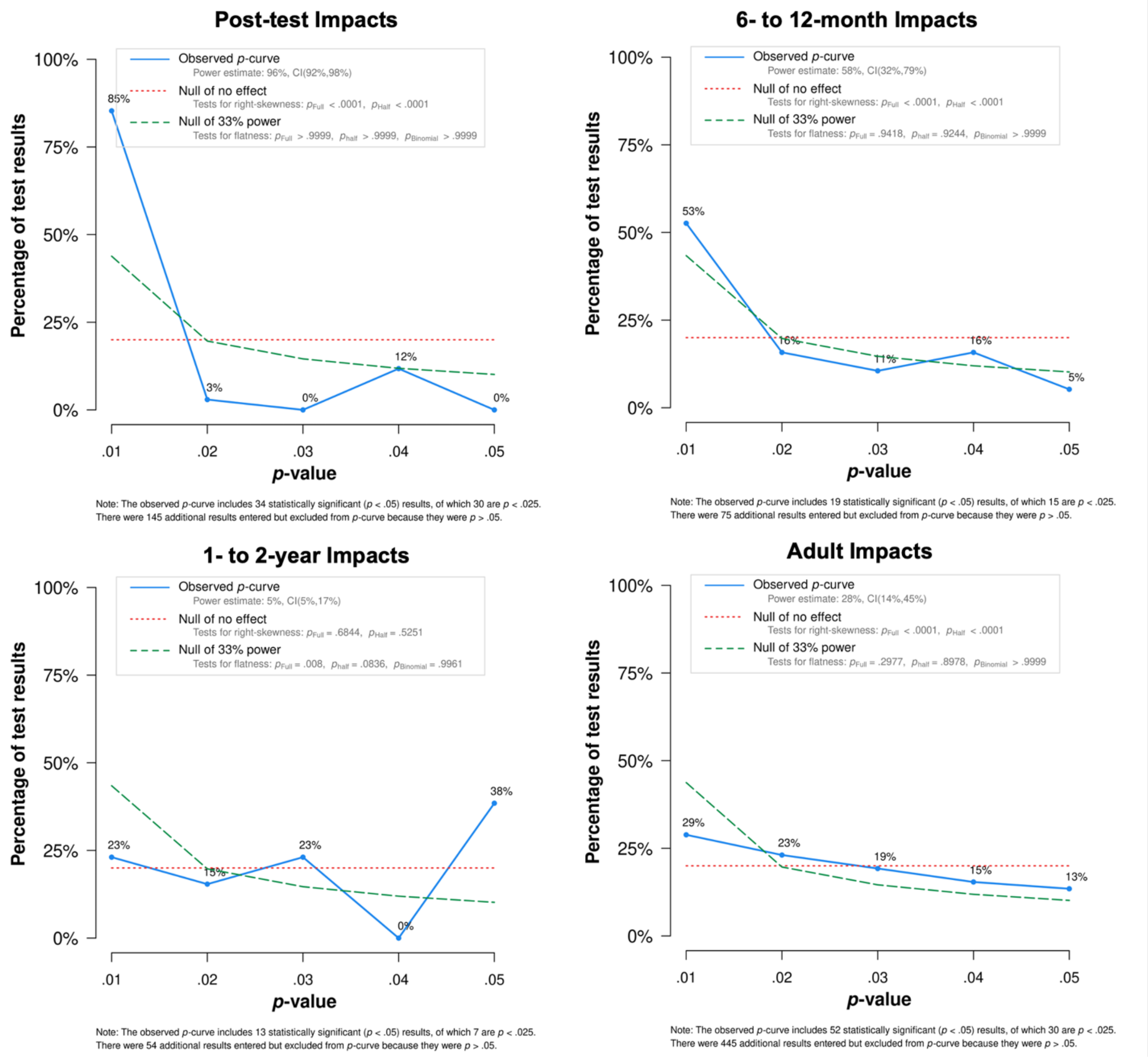
Figure S8
Funnel Plots Using Alternative Standard Errors



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 short-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



Post-test Impacts

Note: The observed *p*-curve includes 34 statistically significant ($p < .05$) results, of which 30 are $p < .025$. There were 145 additional results entered but excluded from *p*-curve because they were $p > .05$.

6- to 12-month Impacts

Note: The observed *p*-curve includes 19 statistically significant ($p < .05$) results, of which 15 are $p < .025$. There were 75 additional results entered but excluded from *p*-curve because they were $p > .05$.

1- to 2-year Impacts

Observed p-curve
Power estimate: 5%, CI(5%,17%)
Null of no effect
Tests for right-skewness: pFull = .6844, pHalf = .5251
Null of 33% power
Tests for flatness: pFull = .008, phalf = .0836, pBinomial = .9961
Percentage of test results
100%
75%
50%
25%
0%
23%
15%
23%
0%
38%
.01
.02
.03
.04
.05
p-value

Note: The observed *p*-curve includes 13 statistically significant ($p < .05$) results, of which 7 are $p < .025$. There were 54 additional results entered but excluded from *p*-curve because they were $p > .05$.

Adult Impacts

Observed p-curve
Power estimate: 28%, CI(14%,45%)
Null of no effect
Tests for right-skewness: pFull < .0001, pHalf < .0001
Null of 33% power
Tests for flatness: pFull = .2977, phalf = .8978, pBinomial > .9999
Percentage of test results
100%
75%
50%
25%
0%
29%
23%
19%
15%
13%
.01
.02
.03
.04
.05
p-value

Note: The observed *p*-curve includes 52 statistically significant ($p < .05$) results, of which 30 are $p < .025$. There were 445 additional results entered but excluded from *p*-curve because they were $p > .05$.

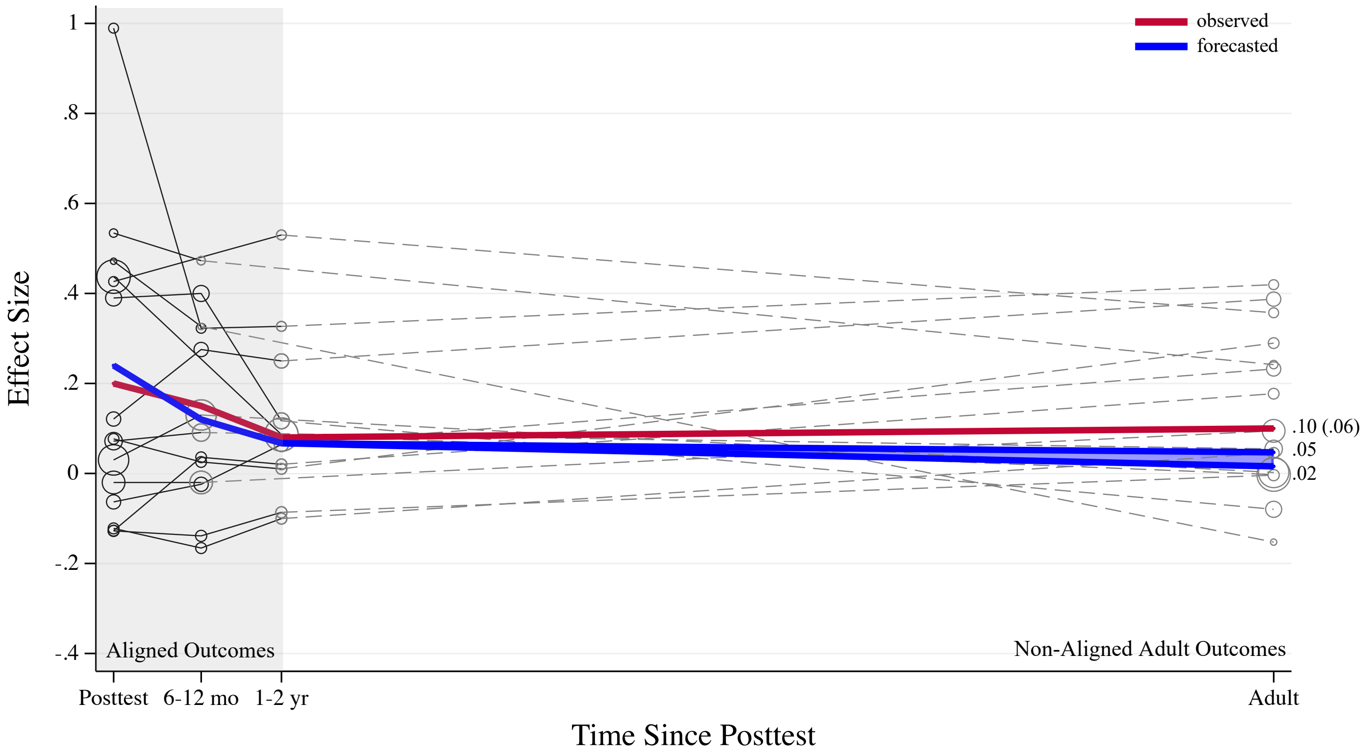
Notes: Each figure presents the *p*-curve for *p*-values of the outcomes contributing effects at each assessment wave. *P*-values were estimated using the alternative standard errors computed through more complex techniques. Figures were created on p-curve.com (Simonsohn, Nelson, & Simmons, 2015). Standard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

Table S6
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 S10
Trajectories of Effects for Interventions with Aligned Short-Term Impacts



Notes: Each line reflects one intervention that measured the same measure and construct at posttest and at least one short-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 2, Panel C for estimates.

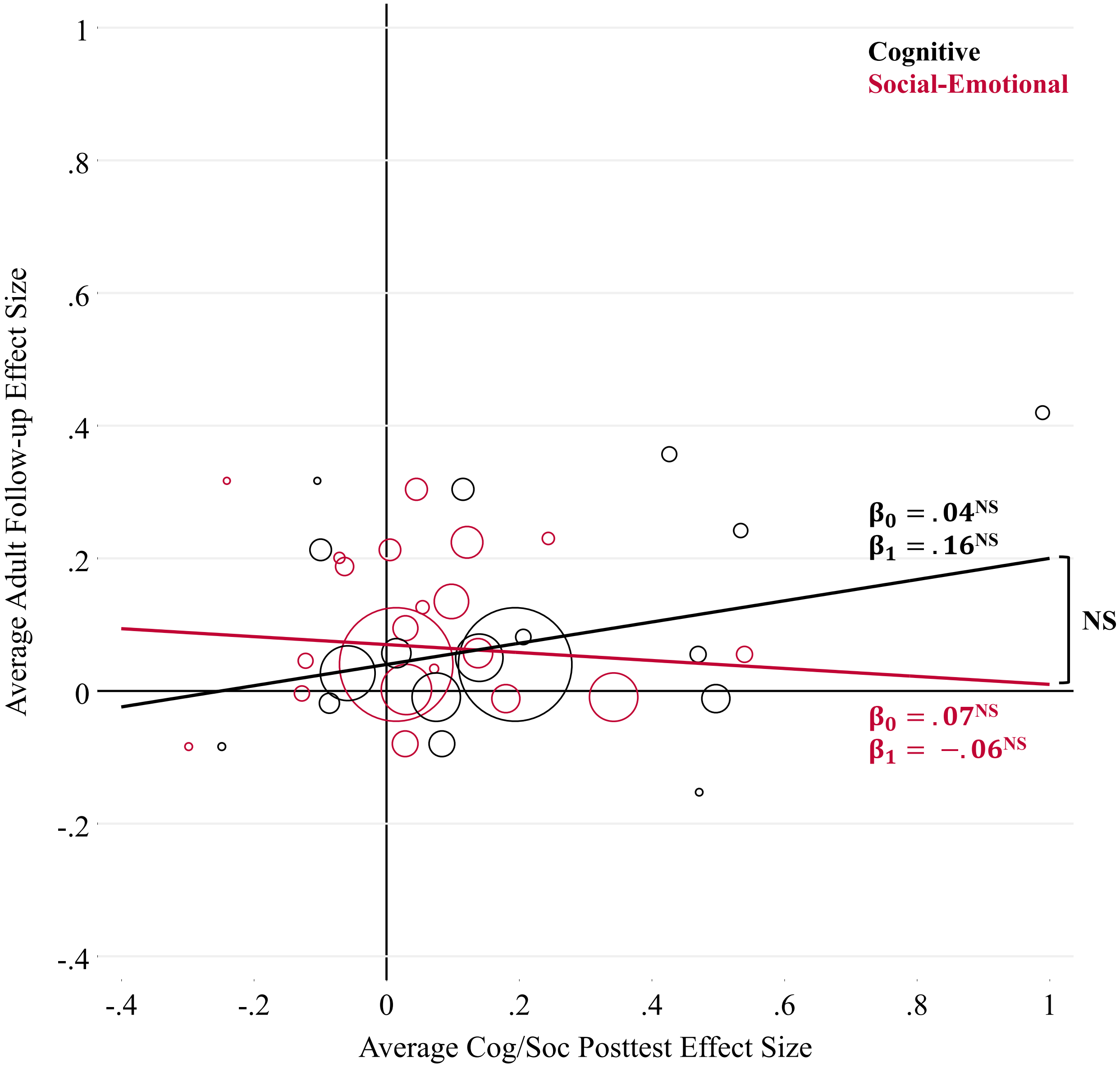
* *p* < .05, ** *p* < .01, *** *p* < .001

Table S7
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 among 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 among 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 S7 for estimates.
* *p* < .05, ** *p* < .01, *** *p* < .001

Table S8

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

	Educational	Substance	Psychological
	(1)	(2)	(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.

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

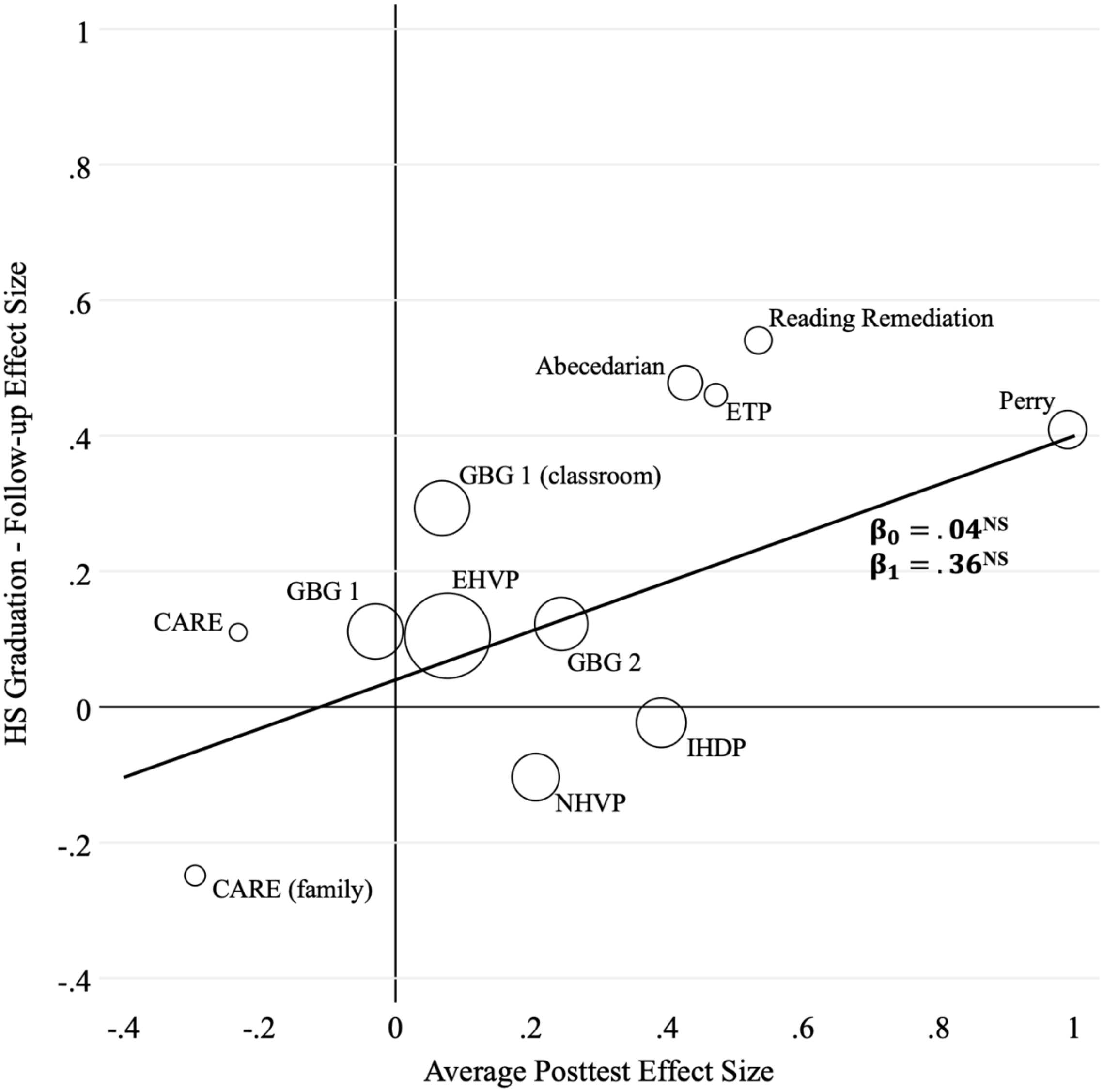
Table S9
Intervention-Level Average Posttest Impacts Predicting High School Attainment

	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.

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

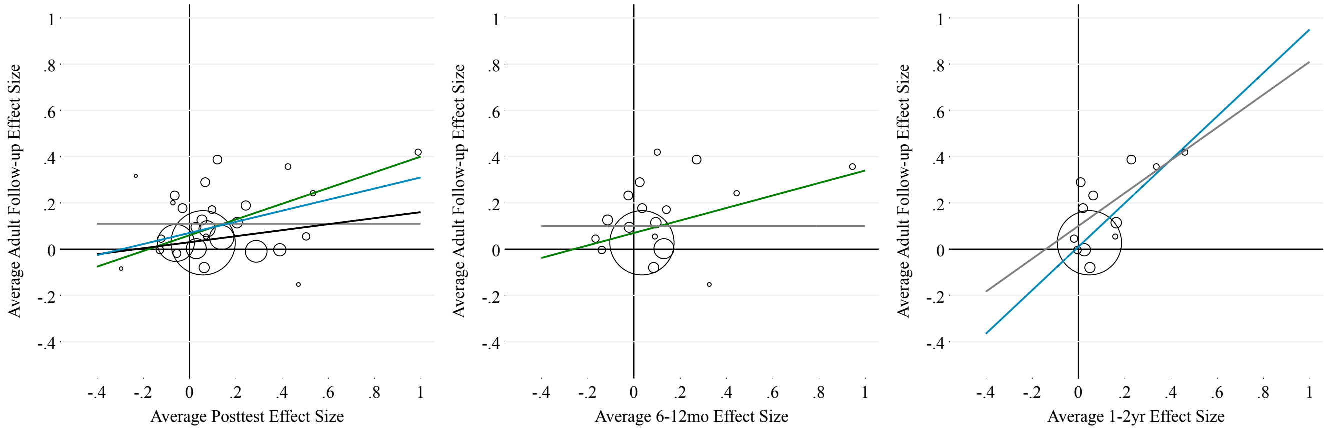
Figure S12
Posttest Impacts Predicting High School Graduation



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 S1 for full intervention names associated with the label abbreviations and Table S9 for estimates.

Figure S13

Posttest and Short-Term Follow-Up Impacts Predicting Adult Impacts



Notes: Figure 1 plots the association between the intervention-level average posttest impact and adult impact. Figures 2 and 3 depict the association between intervention-level average effects at short-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 S2 for estimates.

Table S10
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 S11
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 among posttest and short-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 S12
Probing Study and Data Irregularities

	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: 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, short-term follow-up, and adult impacts and the association among 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.
* $p < .05$, ** $p < .01$, *** $p < .001$

References for Reports Included in the MERF-Emerge 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