ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-782

Do Intervention Impacts on Social-Emotional Skills Persist at Higher Rates than Impacts on Cognitive Skills? A Meta-Analysis of Educational RCTs with Follow-Up

Emma R. Hart
Columbia University

Drew H. Bailey
University of California, Irvine

Sha Luo
Columbia University

Pritha Sengupta
Wesleyan University

Tyler W. Watts
Columbia University

Fadeout is a pervasive phenomenon: post-test impacts on cognitive skills commonly decrease in the years following an educational intervention. Less is known, although much is theorized, about social-emotional skill persistence. The current meta-analysis investigated whether educational RCT impacts on social-emotional skills demonstrated greater persistence than impacts on cognitive skills among 87 interventions involving 59,237 participants and 443 outcomes measured at post-test and at least one follow-up. For post-test impacts of the same magnitude, persistence rates were similar (43% of post-test magnitude) across skill types for follow-ups occurring 6 to 12 months after post-test. At 1- to 2-year follow-ups, persistence rates were larger for cognitive skills (37%) than for social-emotional skills. Interestingly, smaller posttest impacts persisted at proportionately higher rates than larger impacts, which may benefit interventions measuring social-emotional outcomes given their smaller post-test impacts. Considered in whole, social-emotional and cognitive skills demonstrated similar patterns of fadeout.

VERSION: June 2023

Suggested citation: Hart, Emma R., Drew H. Bailey, Sha Luo, Pritha Sengupta, and Tyler W. Watts. (2023). Do Intervention Impacts on Social-Emotional Skills Persist at Higher Rates than Impacts on Cognitive Skills? A Meta-Analysis of Educational RCTs with Follow-Up. (EdWorkingPaper: 23-782). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/7j8s-dy98

Do Intervention Impacts on Social-Emotional Skills Persist at Higher Rates than Impacts on Cognitive Skills?
A Meta-Analysis of Educational RCTs with Follow-up

Emma R. Hart, Drew H. Bailey, Sha Luo, Pritha Sengupta, Tyler W. Watts*

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). We are grateful to Mark Lipsey, Greg Duncan, Jens Dietrichson, Benjamin Lovett, Dana McCoy, and members of the Consortium of Early Childhood Intervention Impact (1R01HD095930-01A1) for their helpful comments on this work. We would like to thank the following research assistants whose efforts made this work possible (in alphabetical order): Helen Ding, Precious Elam, Simran Juneja, Gabby Lammano, Siyu Liang, Sha Luo, Opal Ofstedal, Fatmanur Ozay, Xinyu Pan, Spruha Reddy, Mindy Rosengarten, John Schupbach, Maddie Scricco, Pritha Sengupta, Jessica Sperber, Devon Turner, Leo Weaver, and Josefa Wester.

Abstract

Fadeout is a pervasive phenomenon: post-test impacts on cognitive skills commonly decrease in the years following an educational intervention. Less is known, although much is theorized, about social-emotional skill persistence. The current meta-analysis investigated whether educational RCT impacts on social-emotional skills demonstrated greater persistence than impacts on cognitive skills among 87 interventions involving 59,237 participants and 443 outcomes measured at post-test and at least one follow-up. For post-test impacts of the same magnitude, persistence rates were similar (43% of post-test magnitude) across skill types for follow-ups occurring 6 to 12 months after post-test. At 1- to 2-year follow-ups, persistence rates were larger for cognitive skills (37%) than for social-emotional skills. Interestingly, smaller post-test impacts persisted at proportionately higher rates than larger impacts, which may benefit interventions measuring social-emotional outcomes given their smaller post-test impacts. Considered in whole, social-emotional and cognitive skills demonstrated similar patterns of fadeout.

Keywords: meta-analysis, educational RCTs, fadeout, social-emotional skills

Statement of Relevance

Researchers and policymakers aspire for educational interventions to have long-run impacts on development. In reality, intervention impacts often fade over time. Importantly, most of the research on fadeout has focused on cognitive skills. Researchers have recently argued that concern over fadeout may be overstated and that social-emotional skill impacts may persist in the long term. It could be the case, for example, that interventions that improve children's social-emotional skills initiate developmental cascades that lead to better functioning into adulthood. We tested this possibility in a meta-analytic sample of educational randomized control trials. Overall, we found that, contrary to popular theory, intervention impacts on social-emotional skills did not differ markedly from patterns of impacts on cognitive skills. Instead, intervention impacts faded over time for both skill types. These findings suggest that boosting social-emotional skills may not be a "silver bullet" for generating long-term impacts from educational interventions.

Do Intervention Impacts on Social-Emotional Skills Persist at Higher Rates than Impacts on Cognitive Skills?
A Meta-Analysis of Educational RCTs with Follow-up

Researchers and policymakers often anticipate that educational interventions will improve child outcomes both initially and through cascading long-term impacts. An accumulating body of evidence suggests that initial intervention impacts commonly fade across subsequent follow-up assessments (Bailey et al., 2020). However, most empirical work on fadeout has focused on cognitive skills, making it unclear whether this pattern of diminishing effects exists for social-emotional skills (Abenavoli, 2019). Given arguments that children's social-emotional skills are critical to adult success (Duckworth et al., 2018; Heckman & Kautz, 2012; Nagaoka et al., 2015; Soto et al., 2022) and robust associations between child social-emotional skills and measures of adult success, it is commonly hypothesized that intervention-driven boosts in social-emotional skills will persist more than boosts in cognitive skills.

This hypothesis aligns with the prevailing theory that unmeasured social-emotional skills drive long-term adult impacts of educational programs. Indeed, among a handful of evaluations that have collected adult follow-up, several highly-cited educational interventions have found emerging longer-run impacts on measures of attainment despite observing fadeout, or consistently null effects, on cognitive test scores (e.g., Chetty et al., 2013.; Deming, 2009; Gray-Lobe et al., 2022; for review, see Bailey et al., 2020). In such cases, the persistence of impacts on social-emotional skills, which are commonly unmeasured, has often been inferred to explain how long-term impacts on important life outcomes could be observed despite cognitive skill fadeout (Heckman et al., 2013; Heckman & Kautz, 2012).

Why Might Social-Emotional Impacts Show Greater Persistence?

Two explanations support the possibility that social-emotional skill development systematically differs from cognitive skill development, driving greater impact persistence for social-emotional skills and, ultimately, the emergence of impacts on adult outcomes. The first comes from skill-building models, which assert that more rudimentary skills lay the foundation for advanced skills. Intervention-driven skill boosts may increase the productivity of subsequent skill investments by enabling a child to take advantage of opportunities for growth (i.e., "skills beget skills"; Cunha & Heckman, 2007). Strong social-emotional skills could trigger cascades that reinforce further social-emotional skill development (i.e., "self-productivity") and development in other domains (i.e., "cross-productivity). Although many would predict similar dynamics for cognitive skills, these skill-building theories may be better suited for social-emotional skills, which may receive less instruction in formal educational settings in the absence of the intervention.

Indeed, interventions that effectively boost social-emotional skills could initiate uniquely-effective, positive, *socially-driven* feedback loops between the child and their context, resulting in a developmental trajectory reflective of sustained treatment impacts. For example, an adolescent intervention targeting risky behaviors (e.g., Botvin et al., 2002) may make teens less likely to engage in social drinking. This could, in turn, improve relations with their parents and teachers, leading to further positive reinforcement of prosocial behavior (e.g., see also Social Information Processing Theory; Dodge et al., 1986). Targeting skills at key developmental moments may also produce crucial changes that lead to further skill advancement. For example, reductions in risky behavior could lead teens to circumvent school expulsion, which may otherwise alter long-run trajectories, and continue with low risk-behaviors long-term.

Relatedly, impacts on social-emotional skills may be less prone to fadeout via control group "catch-up." Catch-up occurs when post-treatment contexts and experiences provide children in the control group with opportunities to develop the skills that children in the treatment group acquired from the intervention (Bailey et al., 2020). Control group catch-up has been demonstrated to help explain cognitive skill fadeout (Elango et al., 2015; Watts et al., 2022). Bailey and colleagues (2017) have argued that skill-building interventions are more likely to persist if they target skills unlikely to develop in counterfactual conditions. Whereas many cognitive skills (e.g., math, reading) are explicitly targeted in traditional school settings, social-emotional skills may receive less explicit focus in schools and other learning contexts.

A Skill-Type Null Hypothesis

A skill-based null hypothesis whereby fadeout for social-emotional skills is similar to that for cognitive skills is also possible, as social-emotional skill building may suffer from the same challenges as cognitive skill building. For example, control group catch-up may occur if interventions spur social-emotional development that would have naturally happened through subsequent experiences. Alternately, educational interventions may struggle to initiate skill-building cascades for social-emotional outcomes because treatment impacts are unlikely to overcome the power of the individual-level (e.g., genetics, family environment) and contextual-level (e.g., socioeconomic resources, neighborhood) factors that contribute to the stability of individual differences after interventions end. Indeed, social-emotional capacities demonstrate trait-like stability (Rieger et al., 2017), albeit somewhat less than cognitive skills (Soland et al., 2019[1]).

[1] Factors contributing to stability may also vary by skill type; meta-analytic work has suggested that environmental factors contribute more to stability in personality whereas genetic factors contribute more to cognitive stability (Briley & Tucker-Drob, 2017).

This raises the question: could any process *other* than the persistence of social-emotional skill impacts explain long-run emergent impacts? A more nuanced alternative hypothesis could be that self- and cross-productivity of social-emotional skills, like cognitive skills, are real but limited. It may be that longer-term effects emerge as a product of small carry-over and transfer effects among various skill domains over time rather than as a product of persistence for particular, insular skills. In such cases, impacts on a given skill may diminish, but the intervention could affect long-run outcomes via initially declining treatment impact ripples that spread through complex interconnected network of skills, contexts, and opportunities. The latter include institutional gateways – environmental opportunities influenced by time-specific advantages in social-emotional or cognitive skills that produce more positive long-term outcomes (Bailey et al., 2017, 2020). In the adolescent intervention example, where a boost in social skills prevented expulsion, the short-term gain in social functioning itself may fade, but longer-term effects could emerge as a result of staying in school.

Current Study

The current meta-analysis tested whether educational randomized controlled trial (RCT) impacts demonstrated higher persistence rates for social-emotional outcomes than cognitive outcomes. For this study, we compiled a meta-analytic dataset: the Meta-Analysis of RCTs with Follow-up (MERF). To our knowledge, MERF is the first meta-analysis of educational RCTs that has systematically investigated whether longitudinal impacts unfold distinctly for cognitive and social-emotional skills. MERF was comprised of 86 RCTs with follow-ups on cognitive and/or social-emotional outcomes, sampled from 8 meta-analyses. MERF allowed us to address several methodological issues that have clouded previous work in this area. Notably, to limit internal validity issues we only included RCTs, and to reduce bias due to selective outcome

measurement and reporting across follow-ups we only included constructs consistently measured at post-test and follow-up. For our main examination of the differences in persistence for social-emotional and cognitive skills, we made a priori analytic decisions, but did not have strong a priori hypotheses. The data for this study will be made publicly accessible prior to publication as will the analytic syntax and a detailed meta-analysis protocol detailing all aspects of the dataset creation.

Method

Process

Inclusion Determinations

Interventions were drawn for consideration from eight recent and influential meta-analyses on social-emotional and cognitive intervention effects. These included: Bailey, et al., (2020); Burns et al., (2016); Kraft et al., (2018); Protzko, (2015, 2017); Suggate, (2016); Taylor et al., (2017); Li et al., (2020). Together, these eight meta-analyses provided 426 unique papers, 400 of which had usable PDFs that were reviewed. These 400 papers reported impacts on 298 unique educational interventions. The papers included a diverse array of interventions (e.g., infant home visiting, pre-k, elementary-school-level curriculum, adolescent substance use prevention).

[Figure 1]

Figure 1 presents inclusion decisions. Inclusion criteria and decisions are briefly described below and are further described in the supplemental materials. First, we reviewed each intervention to determine if it utilized an RCT design. Only RCTs were included to limit the need to evaluate the internal validity of quasi-experimental studies, which can be subjective and

difficult to determine. Of the original 298 interventions, 196 utilized an RCT design (102 were excluded). Second, RCTs had to report at least one effect size (or data that could be used to calculate this) for a cognitive or social-emotional outcome to be considered for inclusion. In our sample, 184 RCTs included either cognitive or social-emotional outcomes (12 additional interventions were excluded). Third, interventions had to report follow-up treatment impacts for the same sample of children at least six months after the post-test, with 94 interventions meeting this criterion. (90 interventions were excluded for inadequate follow-up). Fourth, each intervention had to provide usable data (i.e., at least one follow-up effect size or data that could be used to calculate an effect size). Five interventions were removed due to insufficient data, leaving 89 interventions. Finally, although we had not initially excluded studies based on intervention focus, we revisited this decision for four interventions that exclusively focused on nutrition supplementation. These were excluded because they were not educational nor developmental in focus, making them qualitatively dissimilar to the other interventions. Thus, the final sample contained 85 interventions with impacts reported across 139 papers (see supplement for a complete list).

Coding

Papers meeting our inclusion criteria were double-coded for extensive study details and results. The coding team comprised a master coder (the first author, a doctoral student) and two additional coders (Masters-level students). A doctoral-level study PI supervised the coding process. Before coding, the master coder led six months of coding training to ensure coders understood each data element (see supplemental materials for full coding protocol). Coders inputted a variety of information about the intervention, including basic study information (e.g., level of randomization, publication year), intervention and control group details (e.g.,

intervention duration and intensity), treatment targets (e.g., parents, teachers) and inputs (e.g., math skills, self-regulation), internal validity (e.g., whether baseline equivalence was addressed), and participant demographics (e.g., race, sex).

Treatment impacts for cognitive and behavioral outcomes were also coded (e.g., means, standard deviations, effect sizes, *p*-values, etc.), as were details on each reported treatment impact (e.g., author-reported construct name, name of the measure, the timing of assessment). Importantly, we coded pre-test, post-test, and follow-up results. A follow-up wave was included if it reported impacts at least six months after the intervention end. Broad definitions of what constituted cognitive (e.g., IQ, working memory, math, reading, etc.) and behavioral (e.g., behavioral problems, prosocial behaviors, substance use, depressive symptoms, etc.) outcomes were used to guide coding. Additional coding information, and a link to the coding protocol, are provided in the supplement.

After the training period, we tested reliability by checking discrepancies in coding on a random selection of 10 papers in the sample. Across the three coders, agreement ranged from 82% to 89%. All papers were subsequently double-coded. A Masters-level research assistant identified discrepancies in coding, and the coding team frequently met to reach a consensus on all discrepancies. Study PIs were consulted when the coding team could not reach a consensus.

Effect Size and Standard Error Calculations

Following coding, effect sizes were calculated for each outcome on a case-by-case basis. Given that we calculated effect sizes for each outcome available across post-tests and follow-ups, there were multiple effect sizes for a given study. In the most straightforward cases, author-reported effect sizes were used, or effect sizes were calculated using treatment and control group means and standard deviations— details related to these cases are reported below. The

supplement details other crucial aspects of this process, including deviations from typical calculation approaches and calculation formulas (see Figure S1).

When treatment and control group means and standard deviations were reported, effect sizes were calculated using Glass's Delta formula[2]:

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

We used the author-reported effect sizes when they were reported in standardized units, so long as the estimation method produced a viable main effect (e.g., longitudinal effects could not be modeled using parametric assumptions in a growth curve model; the treatment indicator could not be used in interaction terms; mediators could not be included, etc.). There were some cases when viable effect sizes were reported and it was also possible to calculate effects using means and standard deviations. In these cases, we relied on decision criteria to determine which effect size to use in our analyses. The supplement provides details regarding these criteria, but the overarching approach in making effect size determinations was to arrive at the best estimate of the average treatment effect. In all cases, effect sizes were rescaled so that positive effect sizes resulted in "better" outcomes for the treatment group (e.g., a reduction in behavioral problems for the treatment group was rescaled as a positive treatment impact).

Standard errors and *p*-values were also assigned to each effect size. When using author-reported effect sizes, we took the corresponding author-reported standard errors and *p*-values if they were precisely reported. When these statistics were not available, or when effect sizes were calculated using descriptive statistics, standard errors were calculated using:

[2] Standard deviations from the control group were used to calculate effect sizes because the treatment could have impacts on standard deviations.

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

In these cases, *p*-values were estimated by calculating a *t*-statistic from the effect size (i.e., effect size divided by standard error) and determining the associated *p*-value. Degrees of freedom were set to the total sample size minus 2.

Analysis

Analytic Sample

Previous research has identified several methodological factors that may bias meta-analytic estimates of short- and long-run impacts. These factors include selective reporting of short-run impacts, selective reporting of longer-run impacts, and selection into longer-term follow-up in the absence of selective reporting (the first two are addressed in detail in Bailey et al., 2020; Bailey & Weiss, 2022 and Watts et al., 2019 address the third).

To address some of these issues, the analytic sample for the current analysis was created by specifying groupings of constructs measured at post-test and follow-up assessments. We employed a strict operationalization of these groupings for which effect sizes had to meet several criteria. The effect sizes had to come from the same study and experimental group comparison,[3] they had to capture the same author-reported construct, they had to be measured using the same measure and subscale from the same reporter (e.g., self, parent, teacher, etc.), and they had to be collected at post-test and at least one follow-up assessment. Each sequence of effect sizes that met these criteria constituted one unit of analysis in our models. Across interventions, there

[3] Note that there were cases when a study created more than one randomly-assigned treatment group (i.e., the study randomized participants to various different treatment or control conditions).

could be none, one, or many sets of effect sizes depending on how many measures were collected consistently at the post-test and at least one follow-up.

This analytic approach addresses several concerns described above. Additionally, by focusing our main analyses on only measures consistently administered at post-test and follow-up, we are more likely to focus on measures that researchers determined were the key outcomes for a given study. Although we did not explicitly code for whether a given outcome was considered "confirmatory" or "exploratory" by the study authors, measures administered consistently at post-test and follow-up likely constituted the outcomes that the researchers most heavily prioritized. It should be noted that researchers may introduce new and different assessments at subsequent waves for practical (e.g., developmental appropriateness) or theoretical (e.g., selecting new outcomes based on the most promising post-test impacts) reasons that have scientific merit. Thus, we also ran supplemental analyses that relaxed our most restrictive sample inclusion criteria to allow for measures to vary for the same construct across follow-up assessments (see results section for further discussion).

Analytic Models

To answer our key research questions regarding the longitudinal trajectories of cognitive and social-emotional skills, regression analyses were executed to estimate persistence rates for cognitive and social-emotional outcomes separately, and for a combined sample that included both sets of outcomes together. Our modeling approach assumed an underlying causal pathway by which interventions should generate post-test impacts that should predict subsequent follow-up treatment impacts. Thus, we regressed follow-up impacts on post-test impacts to determine the extent to which intervention-driven differences in children's skill persisted when measured at various follow-up assessments. This approach is helpful in that it allows for examination of

relative changes in treatment impacts at follow-up across outcomes that may vary considerably in post-test magnitude. In effect, our approach accounted for changes in intervention impacts over time, considering post-test effect size magnitude.[4]

Outcomes were classified as "cognitive" if they were in these categories: achievement composites, general cognition, language and literacy, math, and other academic abilities. Outcomes were considered "social-emotional" if were in these categories: crime, externalizing behaviors, internalizing symptoms, general social-emotional skills, or substance use. The supplemental file details the construct categorization process, and Table S1 provides examples of constructs and measures for each category.

Random-effects meta-regressions were executed in R using the "metafor" package. The following model was used to estimate patterns of persistence in cognitive versus social-emotional treatment impacts:

Level 1- measure/construct groupings:

$$ES_{fsi} = \beta_{0s} + \beta_1 ES_{psi} + \varepsilon_{fsi}$$

Level 2- study:

$$\beta_{0s} = \gamma_0 + \tau_s$$

[4] Importantly, approaches to characterizing fadeout that rely on examining *absolute* changes in follow-up effects from post-test effects fail to account for the fact that small changes in effects are more or less meaningful at different magnitudes of post-test impact. For example, a reduction in treatment impacts of .05 standard deviations is evidence of substantially more fadeout for an intervention with a post-test impact of .08 standard deviations than a post-test impact of .50 standard deviations. Our regression approach captures *relative* changes in treatment impacts instead of *absolute* changes.

where i indicated analytic group (i.e., the same construct reported at multiple assessment waves collected using the same measure for the same treatment-control group contrast), *s* indicated study, *f* indicated the follow-up assessment wave, and p indicated the post-test assessment wave. Thus, ES_{psi} represents the corresponding effect size for analytic group *i* from study *s* at post-test, and ES_{fsi} represents the corresponding follow-up effect. At level 2, γ_0 captures the constant term for each study. Effects were weighted by $1/se^2$ to place greater weight on effect sizes estimated with greater precision. In these models, follow-up assessments (i.e., the dependent variable) were grouped into three time periods after the intervention ended: at least six months to one year (e.g., six-month follow-up, 12-month follow-up), greater than one year and up to two years (e.g., 14-month follow-up, 24-month follow-up), and greater than two years (e.g., 25-month follow-up, 60-month follow-up).[5]

[Figure 2]

The model estimated the extent to which post-test treatment impacts were predictive of follow-up treatment impacts. This specification provides estimates of two important parameters that are crucial to understanding patterns of fadeout. Figure 2 demonstrates possible patterns for these terms. The first parameter is the slope term (β_1). Here, the slope term captures the predictive association between post-test and follow-up impact, which characterizes the degree to which post-test effects persist at follow-up. In Figure 2, the blue "100% Persistence" line demonstrates a case where post-test effects do not fade (i.e., β_1=1). For example, if this were

[5] Occasionally, there were multiple assessments of the same measure and construct within these categorical time bins. In these cases, estimates were averaged within the following "bins," so that there was one estimate per time bin and measure: six- to twelve-month follow-up, greater than one-year and up to two-years follow-up, greater than two-years and up to three-years follow-up, greater than three-years and up to four-years follow-up, and greater than four-years follow-up. For the purposes of these regression models, the "greater than two years" bin was then reconstructed to capture the longest-term follow-up effect in the case that more than one was available.

true, a post-test effect size of .50 *SD* would predict a follow-up impact of .50, with no fadeout observed. In contrast, the red "0% Persistence" line demonstrates the opposite case: where, regardless of post-test magnitude, the follow-up effect size is 0. Under this scenario, all intervention impacts fadeout regardless of the magnitude of the post-test effect. The green "50% Persistence" line represents a slope of β_1=.50, meaning the follow-up effect would be expected to be 50% of the post-test impact. Here, a post-test impact of .50 *SD* would lead to a follow-up effect of .25 *SD*.

The second key term is the intercept (β_0). This term estimates follow-up impacts when post-test impacts are zero. As such, in this model, the intercept indicates the extent to which factors *other* than post-test effects alone contribute to average follow-up effects. In other words, a non-zero intercept would suggest that follow-up effects are still observed, on average, even when post-test effects are zero. This is demonstrated by the orange line, where a slope of .50 is graphed with a positive y-intercept. Such a pattern of effects might be consistent with the "dark matter" hypothesis that early interventions may affect other latent skills not captured by post-test impacts (e.g., Elango et al., 2015; Pages et al., 2022). Indeed, one could imagine that a positive intercept effect could be observed if an intervention had impacts on an array of skills that drove follow-up effects on the outcome of focus. For example, a broad early childhood intervention could have a zero-post-test impact on mathematics achievement, but one could observe a positive follow-up impact on mathematics ability if the intervention produced impacts on other skills that support mathematical development in later periods (e.g., language). This pattern could also, or additionally, be evidence of measurement error, if meaningful treatment-driven variance in a particular skill is captured at follow-up, but not at post-test.

Likewise, an intercept below 0 would indicate that the follow-up effect is smaller than expected based on the post-test measure, again assuming a linear relation between the post-test and follow-up (see grey dashed line). A negative y-intercept could be expected if the intervention produced long-term adverse effects, even when short-term impacts are positive (e.g., medium-term findings in the Tennessee pre-k study; Lipsey et al., 2018).

With the current data, we tested *Equation 1* on our full set of eligible follow-up impacts across the cognitive and social-emotional domains. We then fit the following model with an interaction term to test our primary question regarding differences between cognitive and social-emotional impacts:

Level 1- measure/construct groupings:

$$ES_{fsi} = \beta_{0s} + \beta_1 ES_{psi} + \beta_2 SOC_{si} + \beta_3 ES_{psi} * SOC_{si} + \varepsilon_{fsi}$$

Level 2- study:

$$\beta_{0s} = \gamma_0 + \tau_s$$

where ES_{fsi} and ES_{psi} are defined as before. Here, we add a dummy indicator, SOC_{si}, capturing whether a construct/measure/treatment-control contrast *i* in study *s* falls within the cognitive or social-emotional category (1= social-emotional outcome; 0 = cognitive outcome). If β_2 were positive, this would indicate that additional, unmeasured factors lead to stronger follow-up effects for cognitive outcomes than social-emotional outcomes (i.e., additional "dark matter").

We then include the interaction between the post-test effect and the social-emotional indicator, denoted by $ES * SOC_{si}$. If β_3 were positive, this would indicate that the persistence rate is greater for social-emotional skills than cognitive skills.

Results

Descriptive Information

Before limiting our sample to post-test and follow-up assessments collected using the same measure, the inclusion process yielded 85 studies with 860 post-test effect sizes and 1,482 follow-up effect sizes. After imposing this limit, the sample contained 68 studies with 87 treatment-control group contrasts, 443 post-test impacts for 59,237 participants, and 572 follow-up impacts. The supplemental file includes forest plots for average post-test cognitive and social-emotional outcomes for these 86 treatment groups (see Figures S2 and S3).

Table 1 details intervention and participant characteristics for these treatment groups and for treatments reporting social-emotional and/or cognitive outcomes specifically. The sample was comprised of papers published from 1969 to 2022 (M = 2005). The majority of these interventions (81%) involved a change in context (i.e., curricular intervention, enhanced Pre-K) rather than the provision of an entirely new environment (i.e., after-school program, Pre-K; 19%). In addition to targeted child outcomes, about 55% of the interventions also targeted teachers and about 23% targeted parents. On average, participants were about 8 years old at baseline, though studies reporting social-emotional outcomes involved older children (M = 11 years old) than those reporting cognitive outcomes (M = 6 years old). Interventions varied considerably in intended treatment length (M = 7 months, range = 1 – 36 months) and, among interventions for which it was possible to compute intended treatment time (only 66% of

interventions), treatment time was higher for cognitive (M = 125 hours) than social-emotional skills (M = 22 hours). This intensity discrepancy was largely driven by a few outlier studies, and our supplemental analyses address the role of age and intervention intensity as potential confounds. Finally, fewer than half of the papers reported characteristics on sample race and ethnicity, making these estimates less instructive.

[Table 1]

Table 2 details the characteristics of the analytic sample split by more granular cognitive and social-emotional outcomes. As Table 2 demonstrates, a total of 54 treatment-control group contrasts (from 40 studies) contributed 238 cognitive constructs measured using the same assessment at post-test and at least one follow-up. On average, we observed between 1 and 2 follow-up assessments for each measure (M = 1.40; range = 1 - 6). The average time elapsed between the post-test and follow-up was 10.37 months. The majority of the cognitive outcomes were language and literacy related (80%), followed by math (9%) and general cognitive outcomes (7%; e.g., IQ). On average, samples were comprised of 367 participants.

[Table 2]

For social-emotional outcomes, 41 treatment-control group contrasts (from 33 studies) contributed 205 effect sizes measured at post-test and follow-up using the same measure. As was the case for cognitive outcomes, we observed between 1 and 2 follow-up assessments for each measure (M = 1.59; range = 1 - 5), and follow-up measures were collected about nine months after the post-test. The most common (43%) social-emotional outcomes were broad assessments of social-emotional skills (e.g., composites of externalizing and internalizing behaviors, prosocial behaviors, etc.) followed by substance use (22%), and internalizing (18%) outcomes. Sample

sizes for social-emotional outcomes were much larger than for cognitive outcomes (average n = 1,420).

Trajectories of Fadeout

Average Effect Sizes Across Assessments

First, we descriptively charted treatment impact trajectories across all follow-up assessments provided. Table 3 presents the average weighted and unweighted effect sizes for all outcomes and for cognitive and social-emotional outcomes considered separately. Figures 3 and 4 display the longitudinal treatment impact trajectories for cognitive and social-emotional outcomes. Across the board, the RCTs had positive impacts at post-test that faded across follow-up. For cognitive and social-emotional outcomes, effects for studies with larger samples hovered closer to null than those with smaller samples.

[Table 3, Figure 3, Figure 4]

This pattern was particularly clear for cognitive outcomes, where we observed a .40 *SD* weighted impact ($p < .001$) that faded to .21 *SD* ($p < .001$) by the 6- to 12-month follow-up, .17 *SD* ($p = .009$) by the 1- to 2- years follow-up, and .05 - .14 *SD* ($p = .08$ - .61) at subsequent follow-ups conducted at least two years after post-test. Social-emotional impacts were smaller at post-test (.14 *SD,* $p < .001$), and minimally different at the 6- to 12-month follow-up (.13 *SD,* $p = .03$). Then, at all subsequent follow-ups, effects were more imprecisely estimated and hovered around zero: -.02 to .08 *SD* ($p = .09$ - .89). Critically, patterns in these descriptives are susceptible to various biases, including selective (non-)reporting of (non-)significant follow-up estimates and selective collection of follow-up data. Thus, to examine patterns of fadeout more rigorously, and with the same outcome matched over time, we turned to our main regression estimates.

Modeled Persistence Rates

After observing these descriptive patterns, we then fit meta-regression models to test the extent to which post-test effects predicted follow-up effects and whether persistence rates differed for cognitive and social-emotional outcomes. Results for 6- to 12-month follow-ups and 1- to 2-year follow-ups can be found in Table 4 and in Figures 5 and 6. Results for estimates greater than two years after post-test exhibited low estimation precision. These results can be found in the supplemental file (see Table S2).

Overall, we found that, compared to cognitive skills, social-emotional skills displayed similar persistence at six- to twelve-month follow-up and less persistence at one- to two-year follow-up. The first parameter of interest, the persistence rate (i.e., slope term), was similar for both cognitive and social-emotional outcomes at the six- to twelve-month follow-up. On average, 6- to 12-month follow-up effects were 43% of the magnitude of post-test effects (β_1 = 0.43, p < 0.001), and the persistence rate did not differ by outcome type (i.e., see interaction in Column 3; p = 0.73). Consistent with these results, the introduction of post-test effect size and outcome type into the model considerably reduced the model heterogeneity (from τ = 0.26 and I^2 = 80% to τ = 0.18 and I^2 = 66%), but the inclusion of the interaction between post-test and outcome type did not (τ = 0.18 and I^2 = 66%). Even with the inclusion of these variables in the model, we still observed considerable unexplained heterogeneity in follow-up effects.

For one- to two-year follow-up, we observed a statistically non-significant overall effect of post-test on follow-up, suggesting little persistence across both skill types (β_1 = -0.03, p = 0.48). This estimate aligned with minimal changes in the explained model heterogeneity with the introduction of post-test effect size and outcome type to the model (τ = 0.15 - 0.17; I^2 = 73% - 74% across models). This suggests that, overall, these variables did not explain variance in

follow-up effects. However, it should be noted that the point estimate for cognitive skills (β_1 = 0.34, $p < 0.001$) was statistically significantly larger than for social-emotional skills (β_1 = -0.08, $p = 0.06$), suggesting a 34% persistence rate for cognitive skills and a near 0% rate for social-emotional skills. Indeed, the inclusion of the interaction between post-test and outcome type reduced model heterogeneity (τ = .08 and I^2 = 73%), though considerable heterogeneity in follow-ups remained. Importantly, these effects were estimated with less precision than the six- to twelve-month findings, given the smaller sample size for this model (n = 89).

[Table 4, Figures 5 & 6]

We did observe evidence of small intercept effects (i.e., "unmeasured mediators"), and small differences by skill type across both models. In Column 1, we observed an intercept across skill types of 0.06 at both of the follow-up timepoints, that was statistically significant at the first follow-up, but not the second. Across both waves, the intercept term for social-emotional skills was statistically significantly larger than for cognitive skills. However, these differences were relatively small (i.e., β_2 ranged from 0.05 to 0.08) and, in the split sample models (see Column 5 and 6), they were not independently statistically significant or apparently substantively different in magnitude, suggesting that the estimates may be imprecise and should be interpreted cautiously.

Exploratory Analyses

Intervention Type

The current analysis focused on charting patterns of persistence across social-emotional and cognitive skill *outcomes.* Although we operated from the assumption that outcomes measured at post-test and at least one follow-up assessment are outcomes of interest that researchers anticipated that their interventions would impact, it is certainly possible that

persistence patterns could vary by intervention focus (i.e., whether the intervention directly targeted social-emotional and/or cognitive skills). Thus, we conducted additional exploratory analyses to determine whether patterns of fadeout varied according to the skills targeted by the intervention: social-emotional skills only, cognitive skills only, or both (i.e., "broad" interventions; see Table 5). In brief, we found no strong or consistent evidence to suggest that the type of intervention heavily influenced persistence rates across follow-up waves. Results should be interpreted cautiously given the small sample sizes for outcome by treatment target subsamples and large standard errors on almost all estimates.

Sensitivity Analyses

Several sensitivity analyses were conducted to test the robustness of the primary findings, and suggested that the findings were generally robust to various model specifications. First, to address the possibility that variation in study-level characteristics could bias estimated patterns of persistence, we tested a model in which we dropped the random effect and, instead, introduced a study-level fixed effect (i.e., the inclusion of a dummy variable control for each study). The inclusion of the study-level fixed effect controlled for any unobserved study-level intervention characteristics and, in effect, constrained our key parameters to studies reporting effects on *both* cognitive and social-emotional outcomes. Overall, this model produced estimates that were substantively aligned with those from the primary model, though less precise (see Table S3, Column 1).

To further examine if key intervention differences could have affected our results, we ran additional models to test whether persistence rates were biased by the two intervention features that notably varied across cognitive and social-emotional impact estimates: participant age and intervention intensity. On average, social-emotional outcomes came from studies with older

participants and fewer hours of intervention content than cognitive outcomes. Thus, we fit a model controlling for average participant age at baseline, which yielded similar estimates to our primary model (Table S3, Column 2). Next, given that intervention intensity was only available for 66% of outcomes, and controlling for this variable would significantly reduce our sample size, we ran models that compared persistence for cognitive and social-emotional outcomes from interventions that were more similar in intervention intensity than was the case in the sample at large. To do so, we limited the sample to interventions involving less than 200 hours of content and less than 100 hours of content. This reduced the average hours of intervention for cognitive outcomes to 49.32 and 32.08 hours, respectively (social-emotional average = 21.73 hours). Interventions that did not report intensity were still included in these two models. Again, these checks produced largely consistent results (Table S3, Columns 3 and 4).

We then turned our attention to clustering issues. Indeed, though all studies in our sample reported treatment impacts at the child level, many studies employed cluster random assignment. Studies that used clustering may vary in key ways from studies that did not. Importantly, the use of clustered randomization (e.g., randomization at the level of classrooms, schools, districts, etc.) inherently precludes larger sample sizes. One might imagine that, given their scale, such studies may also be less intensive and more likely to produce fading effects. This raises concern that our inverse variance weighted estimates erroneously upweight larger-sample studies coming from cluster randomized designs. Recall that, when possible, we took the author-reported impact estimate and standard errors, which presumably included some kind of analytic adjustment for clustering. However, for estimates that we calculated from descriptive tables, it is possible that SEs could contain a downward bias for cluster RCTs. We addressed this concern in several ways. First, we tested a model that used un-weighted treatment impact estimates where studies of

all sample sizes were given equal weight. This model produced estimates that were more inflated, and less precisely estimated than the primary model. As compared to the persistence rate for cognitive skills, the persistence rate for social-emotional skills appeared to be much larger at the 6- to 12-month follow-up, and much smaller at the 1- to 2-year follow-up, though neither of these differences were statistically significant (Table S3, Column 5). We further examined this issue by testing a second model specification that controlled for whether cluster randomization was employed. Primary model estimates did not change with the inclusion of this covariate (Table S3, Column 6). Finally, we tested a more sophisticated model that adjusted the standard errors up for studies that involved clustering to reduce the influence of these studies. Again, estimates were in line with our primary findings (Table S3, Column 7).

Next, we set out to examine the validity of the observed negative post-test impacts. First, we returned to the original papers to check that all negative, statistically significant post-test impacts were correctly coded (see supplement for more discussion). Second, given that theory concerning fadeout generally operates from the assumption that interventions have *positive* post-test impacts, we ran models in which we dropped all outcomes that had negative post-test treatment impacts (approximately 20% of 6- to 12-month follow-up outcomes were dropped, no 1- to 2-year follow-up outcomes were dropped). We again observed results that aligned with our primary estimates (Table S3, Column 8).

Finally, we explored the possibility that effect sizes that required extensive calculation by our team may have biased our results (i.e., effect sizes that had to be determined using calculation assumptions described in the supplement). To address this concern, we fit a model in which we dropped effect sizes that were calculated for dichotomous outcomes, estimated based on imprecise *p*-values, or calculated using estimated standard deviations. These models produced

estimates largely in line with our primary estimates, though the difference between cognitive and social-emotional skill persistence for 1- to 2-year follow-ups was no longer statistically significant (Table S3, Column 9).

Across all of these robustness checks, our primary findings were corroborated. For the 6- to 12-month follow-ups, the sensitivity checks provided estimates that were consistent with estimates from our preferred model. For the 1- to 2-year follow-ups, the robustness check estimates were generally substantively aligned with those from the preferred model, though the magnitude, and statistical significance, of the difference between cognitive and social-emotional skill effects varied model-to-model. Although the effect for cognitive skills remained relatively similar to the preferred model, the magnitude of the social-emotional effect varied some (β_1 = -0.75 to -0.09), indicating that this estimate was sensitive to model specifications.

Publication Bias

On average, about 42% of post-test effects were statistically significant at $p < .05$. Figure S4 displays funnel plots for the post-test and follow-up effects included in the analyses, averaged within treatment-control group contrast. To statistically test for evidence of publication bias, we conducted a PEESE test. First, we ran our primary null model with the inclusion of standard errors as a predictor.[6] Consistent with the possibility of publication bias, larger standard errors were predictive of larger effect size magnitudes, particularly at post-test (β_1 = 22.19; $p < .0001$), but also at 6- to 12-month follow-up (β_1 = 2.34; $p < .0001$), 1- to 2-year follow-up (β_1 = .97; p = .05), and greater than 2-year follow-up (β_1 = .71; p = .07). However, there are other plausible explanations for an effect size-precision association. For example, interventions from smaller

[6] In other words, we tested a random-effects meta-regression model predicting effect size at post-test, 6- to 12-month follow-up, 1- to 2-year follow-up, and greater than 2-year follow-up, with no independent variables or moderators other than a control for standard errors.

samples may have been more intensive and/or targeted and thus produced larger impacts. We then tested whether the inclusion of the standard errors as a covariate in our primary models changed estimated patterns of persistence. Our results were substantively aligned with the estimates from our primary model (Table S4).

To further probe the possibility of publication bias, we examined the distribution of *p*-values. Figure S5 presents the relative frequency of *p*-values statistically significant at $p < .05$ for each follow-up assessment wave. The distribution of *p*-values for effect sizes reported at post-test and 6- to 12-month follow-up provided little evidence of *p*-hacking. However, at later follow-ups greater than one year after post-test, there appeared to be an uptick in *p*-values close to .05, suggesting a greater proportion of estimated impacts may be inflated relative to population values.

To avoid estimate inflation due to publication- and reporting-related biases, our primary estimates were generated using models that required the same construct be measured at post-test and at least one follow-up using the same measure at each assessment wave. Indeed, this approach is limited in that some studies may have legitimate reasons, such as changes in participant age, that require the use of different measures across follow-up assessments. Thus, we ran an alternate model in which we allowed the same construct to be assessed using different measures at post-test and follow-ups. The persistence rates were relatively similar to those from the preferred aggregation approach (Table S5, Column 1).

To check our assumption that using a broad aggregation approach, such as that employed in other studies, would inflate our estimates, we ran a model in which we averaged all cognitive or social-emotional outcomes at post-test and each follow-up assessment, respectively. When the data were aggregated without consideration for measure type or construct alignment (Table S5,

Column 2), the persistence rates were inflated, as expected, demonstrating greater social-emotional persistence at the six- to twelve-month follow-up and the opposite pattern at the one- to two-year follow-up.

Selection into Longer-Term Follow-up

One concern in interpreting our findings is that follow-up assessments may be disproportionately collected for outcomes that showed promising treatment impacts at post-test. Indeed, given limited bandwidth and resources, researchers may opt to collect follow-up assessments for outcomes with larger post-test effects, and grant-funding agencies may only fund follow-ups when post-test impacts are substantive. This selection bias could affect our estimates by limiting the available population of follow-up effects to only studies that showed sizeable post-test impacts. To evaluate the likelihood of this possibility, we checked the post-test effect sizes associated with the sample of follow-up effects reported at each assessment wave (see Table 3).

We found little evidence for selection into follow-up up to 2 years after post-test. However, post-test effect sizes at later follow-up waves were generally larger than initial post-test effect sizes. This pattern was generally observed for unweighted post-test estimates but not for weighted post-test estimates. Given the down-weighting of small-sample studies in the weighted estimates, this pattern suggests that selection into follow-up may be a more significant issue for smaller-sample studies, which were more common among cognitive outcomes. This pattern may, indeed, be indicative of bias in the reporting and collection of long-run cognitive effects and in long-term data collection decisions. This also suggests that our estimates of average follow-up impacts (see Table 3) may be inflated, especially those reported greater than two-years after post-test.

Discussion

The current study investigated the theory that educational intervention impacts on social-emotional skills persist more than impacts on cognitive skills. We used a meta-analytic approach to examine data from educational RCTs that reported post-test and follow-up intervention impacts on the same cognitive and social-emotional outcomes using the same measures overtime. Overall, we observed evidence of fadeout: short-run follow-up effects were 43% of the magnitude of immediate post-test effects and became less predictive at subsequent follow-up waves. Our analyses demonstrated robust evidence for similar persistence rates for cognitive and social-emotional skills at six- to twelve-month follow-up, and greater persistence rates for cognitive skills at one- to two-year follow-up. Additionally, we observed unmeasured mediator effects for both outcomes, and larger effects for social-emotional outcomes. We found meaningful heterogeneity in follow-up effects that was unexplained by post-test treatment impacts. These results suggest theories purporting that interventions targeting social-emotional skills will generate more persistent effects require revision.

Interestingly, our results suggest that persistence rates may vary by post-test effect magnitude. For post-test impacts of the *same magnitude*, persistence rates are similar for cognitive and social-emotional skills at 6- to 12-months follow-up, and more favorable for cognitive skills at 1- to 2-years follow-up. Given the presence of small intercept effects for both skill types, when post-test impacts are small, or even zero, our results suggest that there is proportionally stronger persistence than when impacts are larger. This finding may arise because interventions produce positive effects on unmeasured skills that transfer to later measured skills. This could be why average long-term impacts for both skill types appear to asymptote above zero (though such effects are not usually statistically distinguishable from zero). Interestingly,

because social-emotional outcomes had smaller post-test impacts (and possibly larger intercept effects), social-emotional impacts may persist relatively *more* when compared with larger post-test impacts on cognitive skills, despite similar patterns of persistence for impacts of the same magnitude. However, our models suggest that such longer-term impacts are likely to be small in absolute magnitude.

Still, the current study did not find evidence to suggest that an intervention would have more persistent effects on targeted skills if it produced a 1 *SD* gain on social-emotional skills compared to an intervention that produced a 1 *SD* gain on cognitive skills. The similar persistence rate for both skill types suggests that social-emotional skills may be susceptible to the mechanisms that drive cognitive skill fadeout. For example, intervention-targeted social-emotional skills may be likely to naturally develop in subsequent contexts facilitating control group catch-up. Educational interventions may also show diminishing impacts on social-emotional skills because they do not overcome the influence of many individual, contextual, and societal factors that continue to shape skill development when interventions end (see Watts et al., 2017). Given that we observed substantial heterogeneity in follow-up impacts that was unexplained by post-test impacts, it may be the case that traditional skill-building models do not reflect the complexities of skill development.

Interventionists might view this news negatively or positively. On the one hand, it is disappointing that we have yet to identify a large class of skills for which end-of-treatment impacts persist at the same magnitude indefinitely. On the other hand, given many previous findings of positive long-term impacts of educational interventions on important adult outcomes, mediating processes must exist and plausibly include both cognitive and social-emotional skills, despite diminishing impacts after treatment. Although impacts on a focal skill may diminish,

long-run impacts may develop through small impacts on a network of unmeasured complementary skills, contexts, and opportunities, with impacts stabilizing at some non-zero (but statistically undetectable) level that is lower than initial impacts. In light of this, predicting a priori which skill impacts will show the most persistence may be difficult if persistence is contingent on complex interactions between the child and their environment.

Several limitations are worth noting. Critically, despite including a relatively large collection of RCT studies in the meta-analytic sample, there was limited statistical power to estimate the relation to follow-up impact persistence greater than two years after post-test. This imprecision precludes concrete conclusions about the relation between post-test and longer-run follow-up impacts. This problem is partly a symptom of a larger issue of limited grant funding for collecting long-term follow-up data for educational RCTs, and that funding is often allocated to RCTs that demonstrate large post-test impacts (Watts et al., 2019). Second, it is important to note the generalizability of our findings. 80% of cognitive outcomes in our models capture language or literacy outcomes. The social-emotional constructs were more diverse although these were largely survey-based, which reflects the state of the field. Indeed, it is possible that measures of socio-emotional development are not adept at capturing the sorts of "soft-skill" gains many interventions hope to promote.

These findings suggest that social-emotional skills may not be the single class of missing mediators of emergent long-run impacts on adult outcomes. Emergence could be consistent with other explanations, such as persistent-but-hard-to-measure social-emotional *and* cognitive impacts (Reynolds & Ou, 2011), and/or impacts on institutional gateways that can generate longer-term benefits even in the absence of medium-run skill impacts (see Bailey et al., 2020; Pages et al., 2022). In light of considerable heterogeneity in follow-up effects that was

unexplained by post-test impacts, future work will investigate theoretically-motivated intervention- and participant-level characteristics as moderators of persistence, and explore the extent to which these moderators and post-test impacts forecast the emergence of long-run outcomes.

References

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., Jenkins, J. M., & Alvarez-Vargas, D. (2020). Complementarities between early educational intervention and later educational quality? A systematic review of the sustaining environments hypothesis. *Developmental Review*, *56*, 100910. https://doi.org/10.1016/j.dr.2020.100910

Bailey, D., & Weiss, M. J. (2022, April). Do meta-analyses oversell the longer-term effects of programs? (Part 1). MDRC. https://www.mdrc.org/publication/do-meta-analyses-oversell-longer-term-effects-programs-part-1#:~:text=The%20most%20obvious%20implication%20of,intervention%20on%20longer%2Dterm%20outcomes.

Botvin, G., Griffin, K., Nichols, T., & Ifill-Williams, M. (2002). Preventing binge drinking during early adolescence: One and two-year follow-up of a school-based preventive intervention. *Psychology of Addictive Behaviors : Journal of the Society of Psychologists in Addictive Behaviors*, *15*, 360–365. https://doi.org/10.1037/0893-164X.15.4.360

Briley, D. A., & Tucker-Drob, E. M. (2017). Comparing the developmental genetics of cognition and personality over the life span. *Journal of Personality*, *85*(1), 51–64. https://doi.org/10.1111/jopy.12186

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

Campbell, F. A., Pungello, E. P., Burchinal, M., Kainz, K., Pan, Y., Wasik, B. H., Barbarin, O. A., Sparling, J. J., & Ramey, C. T. (2012). Adult outcomes as a function of an early childhood educational program: An Abecedarian Project follow-up. *Developmental Psychology*, *48*, 1033–1043. https://doi.org/10.1037/a0026644

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. American economic review, 104(9), 2633-2679.

Cunha, F., & Heckman, J. (2007). The Technology of Skill Formation. *The American Economic Review*, *97*(2), 31–47.

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

Deming, D. J. (2017). The growing importance of social skills in the labor market*. *The Quarterly Journal of Economics*, *132*(4), 1593–1640. https://doi.org/10.1093/qje/qjx022

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

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

Elango, S. , Heckman, J. J., García, J. L., & Hojman, A. (2015). Early childhood education. *NBER Working Paper 21766.*

Goldhaber, D., Jin, Z., & Startz, R. (2022). How much do early teachers matter? Working Paper No. 264-0422. *National Center for Analysis of Longitudinal Data in Education Research (CALDER)*.

Gray-Lobe, G., Pathak, P. A., & Walters, C. R. (2022). 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

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., 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.

Li, W., Duncan, G. J., Magnuson, K., Schindler, H. S., Yoshikawa, H., & Leak, J. (2020). Timing in early childhood education: How cognitive and achievement program impacts vary by starting age, program duration, and time since the end of the program. Annenberg Institute for School Reform at Brown University, (20-201).

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547–588. https://doi.org/10.3102/0034654318759268

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

Masten, A., Roisman, G., Long, J., Burt, K., Obradović, J., Riley, J., Boelcke-Stennes, K., & Tellegen, A. (2005). Developmental cascades: Linking academic achievement and externalizing and internalizing symptoms over 20 years. *Developmental Psychology*, *41*, 733–746. https://doi.org/10.1037/0012-1649.41.5.733

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

Nagaoka, J., Farrington, C. A., Ehrlich, S. B., Heath, R. D., Johnson D. W., Dickson, S., Tuner, A. C., Mayo, A., & Hayes, K. (2015). Foundations for young adult success: A developmental framework.

Pages, R., Bailey, D. H., & Duncan, G. J. (2022). Exploring the "dark matter" of early childhood educational programs: A pattern-of-indirect-effect approach.

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

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

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

Rieger, S., Göllner, R., Spengler, M., Trautwein, U., Nagengast, B., & Roberts, B. W. (2017). Social cognitive constructs are just as stable as the Big Five between Grades 5 and 8. *AERA Open*, *3*(3), 2332858417717691. https://doi.org/10.1177/2332858417717691

Soland, J., Kuhfeld, M., Wolk, E., & Bi, S. (2019). Examining the state-trait composition of social-emotional learning constructs: Implications for practice, policy, and evaluation. *Journal of Research on Educational Effectiveness*, *12*(3), 550–577. https://doi.org/10.1080/19345747.2019.1615158

Soto, C. J., Napolitano, C. M., Sewell, M. N., Yoon, H. J., & Roberts, B. W. (2022). Going beyond traits: Social, emotional, and behavioral skills matter for adolescents' success. *Social Psychological and Personality Science*, 19485506221127484. https://doi.org/10.1177/19485506221127483

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

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child Development*, *88*(4), 1156–1171. https://doi.org/10.1111/cdev.12864

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., Bailey, D. H., & Mattera, S. (2022, March). Exploring Patterns of Fadeout across Multi-Site RCTs of an Early Childhood Curriculum Intervention. In 2021 APPAM Fall Research Conference. APPAM.

Watts, T. W., Clements, D. H., Sarama, J., Wolfe, C. B., Spitler, M. E., & Bailey, D. H. (2017). Does Early Mathematics Intervention Change the Processes Underlying Children's Learning?. *Journal of Research on Educational Effectiveness*, *10*(1), 96-115. https://doi.org/ 10.1080/19345747.2016.1204640

Figure 1
Inclusion/Exclusion Flow

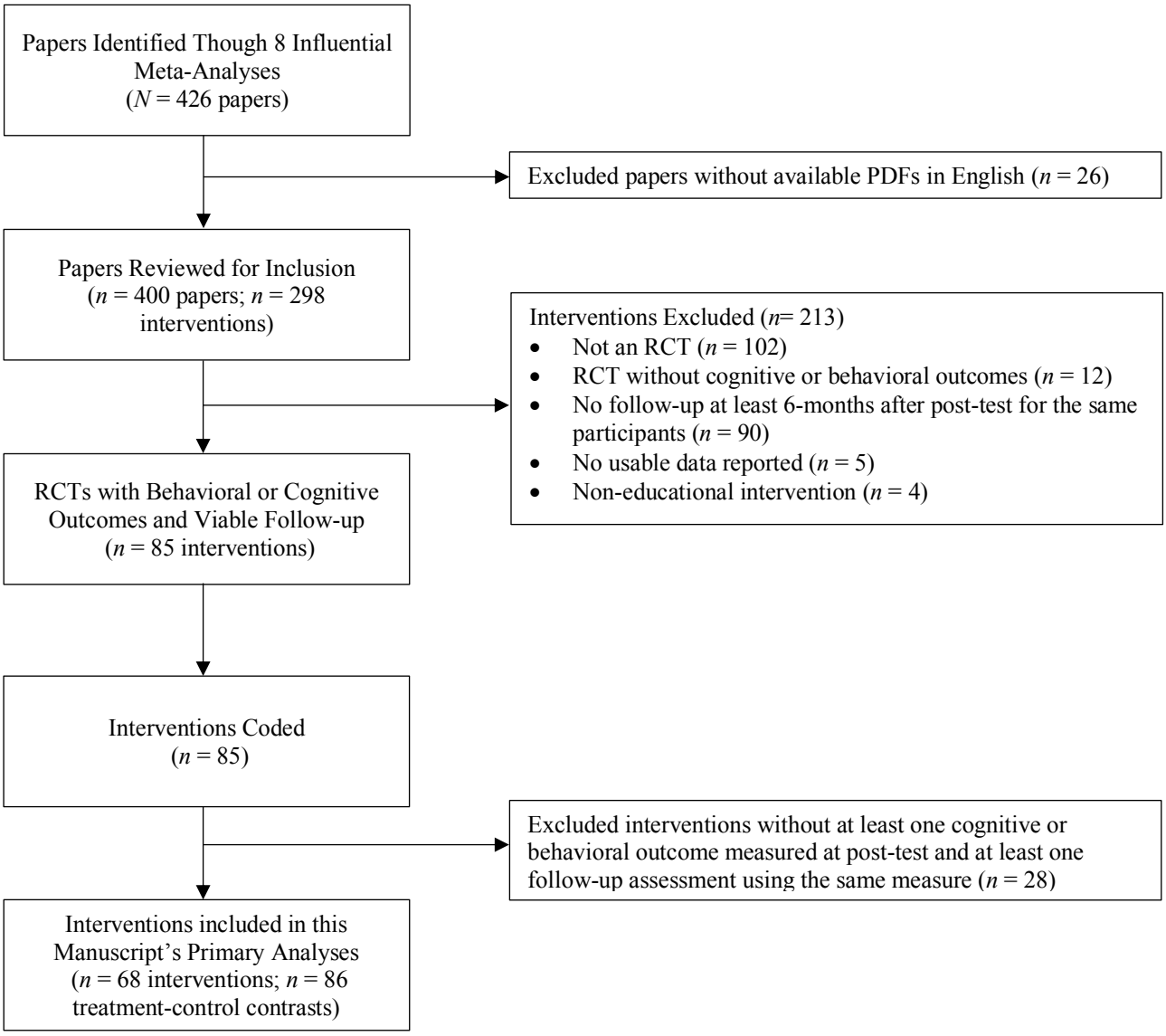


Figure 2
Hypothetical Patterns of Fadeout/Persistence

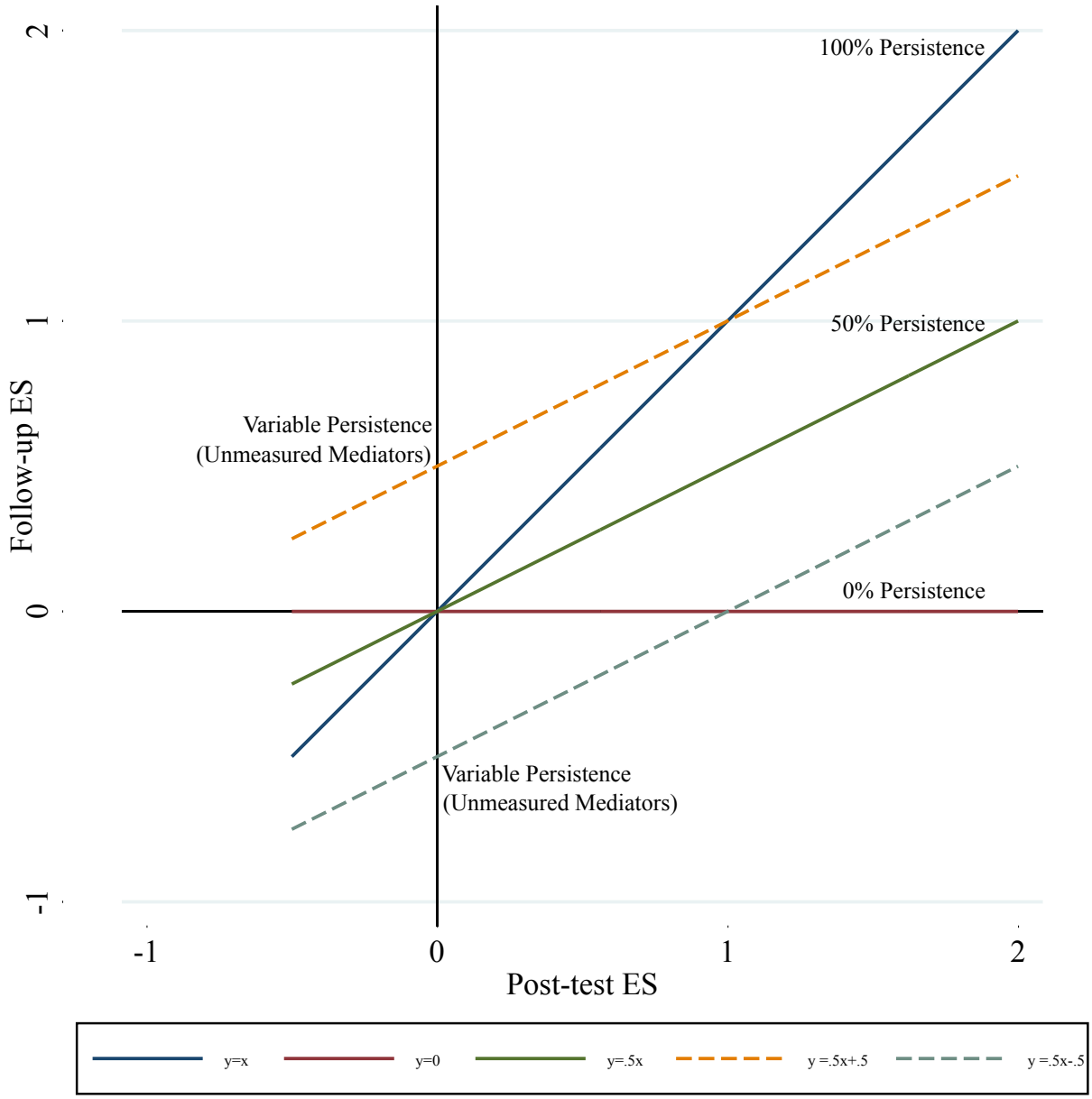


Table 1
Intervention and Participant Characteristics (mean [minimum- maximum])

	All Outcomes (1)	n	Social-Emotional Outcomes (2)	n	Cognitive Outcomes (3)	n
Paper Publication Year	2005 [1969-2022]	86 (100%)	2006 [1987-2015]	41 (100%)	2005 [1969-2022]	54 (100%)
Baseline Age (months)	95.17 [0-170]	85 (99%)	126.22 [40.74-170.00]	40 (98%)	67.81 [0.00-122.00]	54 (100%)
Intervention Length (months)	7.23 [0.92-36.00]	74 (86%)	6.34 [0.92-18.00]	31 (76%)	7.92 [0.92-36.00]	49 (91%)
Intervention Time (hours)	88.33 [4.00-1075.39]	57 (66%)	21.73 [6.00-162.94]	20 (49%)	125.34 [4.00-1075.39]	38 (70%)
Intervention Type (%)						
Change in Environment	80.56	72 (83%)	87.88	33 (80%)	76.60	47 (87%)
New Environment	19.44	72 (83%)	12.12	33 (80%)	23.40	47 (87%)
Intervention Targets (%)						
Math	6.90	87 (100%)	7.32	41 (100%)	11.11	54 (100%)
Language/Literacy	51.72	87 (100%)	9.76	41 (100%)	83.33	54 (100%)
Science	1.15	87 (100%)	0.00	41 (100%)	1.85	54 (100%)
General Cognition	5.75	87 (100%)	2.44	41 (100%)	9.26	54 (100%)
Executive Functioning	1.15	87 (100%)	0.00	41 (100%)	1.85	54 (100%)
Learning Skills	1.15	87 (100%)	0.00	41 (100%)	1.85	54 (100%)
Social-Emotional Skills	52.87	87 (100%)	100.00	41 (100%)	24.07	54 (100%)
Substance Use	13.79	87 (100%)	29.27	41 (100%)	0.00	54 (100%)
Psychological Wellbeing	11.49	87 (100%)	24.39	41 (100%)	0.00	54 (100%)
Cognitive Only	42.53	87 (100%)	90.24	41 (100%)	7.41	54 (100%)
Social-Emotional Only	44.83	87 (100%)	0.00	41 (100%)	72.22	54 (100%)
Broad	12.64	87 (100%)	9.76	41 (100%)	20.37	54 (100%)
Adult Involvement (%)						
Teachers	55.17	87 (100%)	60.98	41 (100%)	55.56	54 (100%)
Parents	22.99	87 (100%)	26.83	41 (100%)	24.07	54 (100%)
Participant Race/Ethnicity (%)						
Asian	13.99 [2.00-45.07]	15 (17%)	10.46 [2.00-43.72]	6 (15%)	19.08 [5.13-45.02]	10 (19%)
Black	41.12 [2.00-100.00]	42 (48%)	33.52 [2.00-97.00]	20 (49%)	46.99 [7.60-100.00]	30 (56%)
White	55.75 [2.00-98.28]	37 (43%)	62.24 [3.00-98.28]	21 (51%)	42.85 [2.00-80.00]	22 (41%)
Hispanic	24.43 [1.90-100.00]	29 (33%)	16.31 [1.90-43.72]	15 (37%)	33.08 [4.90-100.00]	20 (37%)
Female Participants (%)	47.40 [28.00-100.00]	73 (84%)	50.84 [40.74-100.00]	37 (90%)	44.85 [28.00-54.00]	44 (81%)
Average n (at post-test)	737.68 [24-10170]	87 (100%)	1334.43 [42-10170]	41 (100%)	378.46 [24-3929]	54 (100%)

Note: This table presents intervention and participant characteristics for treatment-control group contrasts that contributed "aligned groups" to the analytic sample. Column 1 presents these characteristics for all outcomes contributing analytic groupings to the analytic sample, whereas Columns 2 and 3 present characteristics contributing at least one social-emotional or cognitive analytic group, respectively. "N" indicates the number of treatment-control contrasts that reported information on the characteristic and, what percentage of treatment-control group contrasts contributed to the averages (i.e., how representative the averages are of the full analytic sample). Averages are presented with accompanying ranges (minimum and maximum) when appropriate.

Table 2
Analytic Sample Characteristics for Cognitive and Social-Emotional Outcomes

	ESs from group (%)	Treatment Groups (#)	Aligned Groupings (#)	Avg # of Follow-ups	Avg Months since Post-test	Avg Post-test ES, weighted	Avg Post-test ES (SE), unweighted	Post-test ES p <.05	Avg N
Cognitive Outcomes		**54**	**238**	**1.40**	**10.37**	**0.40 (0.06)*** **	**0.39 (0.06)*** **	**33%**	**367**
Language and Literacy	80%	46	197	1.29	9.27	0.44 (0.07)***	0.43 (0.07)***	36%	310
Math	9%	18	22	1.35	11.58	0.07 (0.08)	0.03 (0.21)	12%	608
Cognitive- General	7%	7	12	2.54	20.87	0.39 (0.12)**	0.42 (0.29)	41%	225
Other Academic Ability	2%	4	5	1.00	5.20	0.12 (0.03)***	0.09 (0.45)	10%	794
Achievement Composite	2%	2	2	3.25	19.50	0.25 (0.06)***	0.26 (0.71)	25%	2068
Social-Emotional Outcomes		**41**	**205**	**1.59**	**8.74**	**0.14 (0.04)*** **	**0.17 (0.07)***	**34%**	**1420**
Social-Emotional- General	43%	28	89	1.57	8.39	0.17 (0.06)**	0.21 (0.11)*	32%	1471
Substance Use	22%	12	53	1.00	6.63	0.13 (0.06)*	0.18 (0.14)	56%	2596
Internalizing	18%	17	34	1.69	7.81	0.10 (0.06)+	0.12 (0.17)	23%	565
Externalizing	14%	8	25	2.12	12.32	0.09 (0.03)**	0.08 (0.2)	25%	610
Criminality	3%	2	4	3.00	16.95	--	0.07 (0.5)	19%	883
Treatment Type									
Cognitive Only TXs	44%	39	195	1.19	8.74	0.44 (0.07)***	0.40 (0.07)***	34%	119
Social-Emotional Only TXs	45%	37	184	1.44	7.79	0.14 (0.04)**	0.17 (0.07)*	34%	1449
Broad TXs	11%	11	51	2.46	17.21	0.35 (0.1)***	0.28 (0.14)*	33%	1201

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

Note: "ES" = Effect size. Effect sizes are in standard deviation units. Weighted effect sizes are weighted by $\frac{1}{se^2}$. Average number of assessments reflects the average number of follow-up assessments that were collected (at least 6-months after post-test). Timing of assessments refers to the average number of months that elapsed between post-test at follow-up assessment(s). The number of aligned groups refers to the number of groupings that included a post-test and at least one follow-up assessment of the same construct measured using the same measure, subscales, and reporter within a treatment-control contrast.

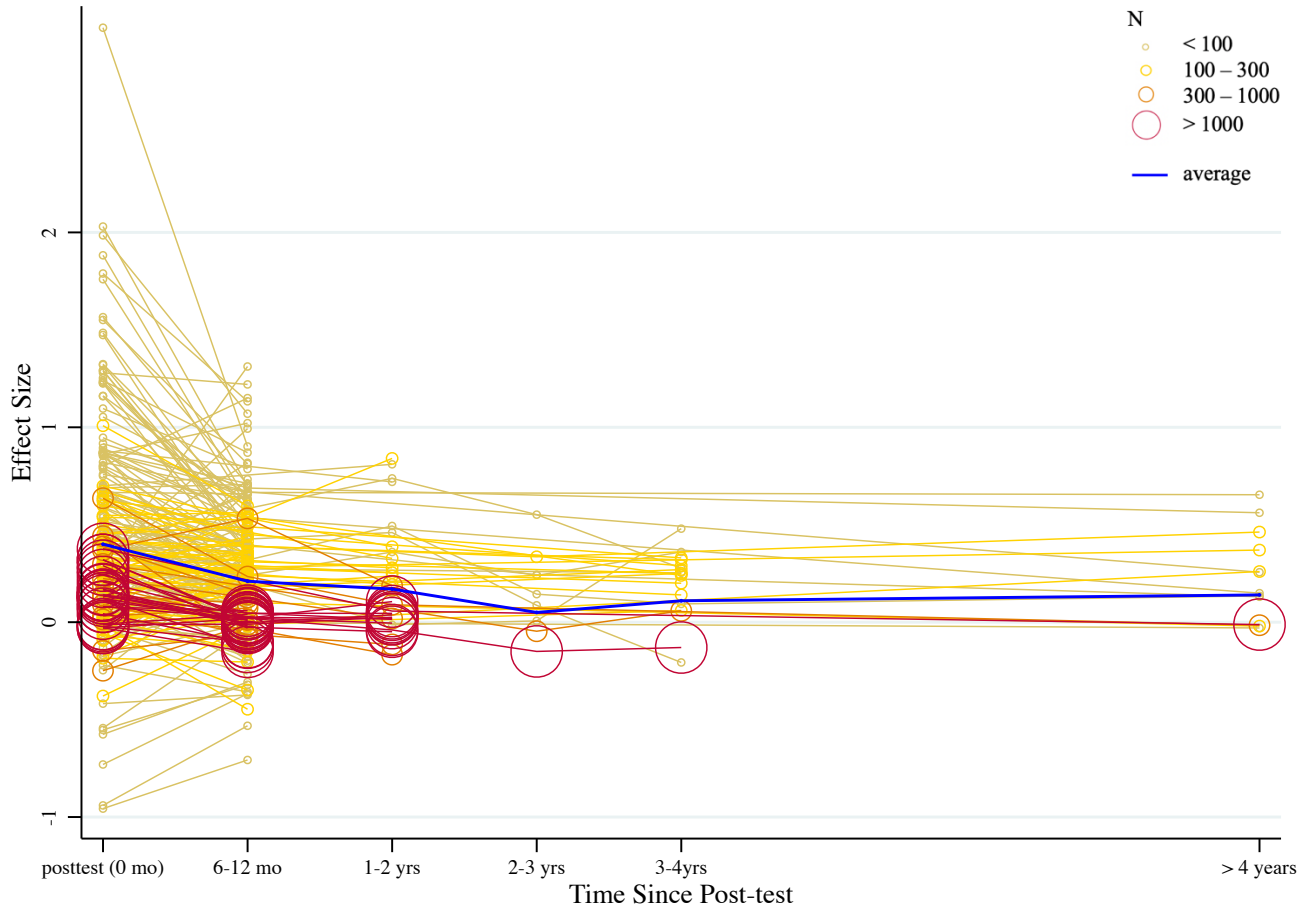
Table 3
Average Effect Sizes Across Post-test and Follow-up Assessment Waves

	All Outcomes			Social-Emotional Outcomes			Cognitive Outcomes		
	Avg. ES (SE) /wave (1)	Avg. Post-test ES (SE) for obs. /wave (2)	n	Avg. ES (SE) /wave (3)	Avg. Post-test ES (SE) for obs. /wave (4)	n	Avg. ES (SE) /wave (5)	Avg. Post-test ES (SE) for obs. /wave (6)	n
Panel A: Weighted Average Effect Sizes									
Post-test	0.29 (0.04)***	0.29 (0.04)***	443	0.14 (0.04)***	0.14 (0.04)***	205	0.40 (0.06)***	0.40 (0.06)***	238
6 months to 1 year	0.19 (0.04)***	0.3 (0.05)***	413	0.13 (0.06)*	0.15 (0.04)***	188	0.21 (0.04)***	0.41 (0.06)***	225
> 1 year, up to 2 years	0.06 (0.04)	0.24 (0.06)***	89	0.00 (0.04)	0.08 (0.04)+	59	0.17 (0.06)**	0.44 (0.09)***	30
> 2 years, up to 3 years	0.02 (0.07)	0.25 (0.09)**	31	-0.02 (0.11)	0.10 (0.03)**	22	0.05 (0.1)	0.52 (0.16)**	9
> 3 years, up to 4 years	0.09 (0.04)*	0.33 (0.11)**	26	0.08 (0.04)+	0.10 (0.10)	8	0.11 (0.08)	0.47 (0.13)***	18
> 4 years	0.14 (0.08)+	0.49 (0.08)***	13				0.14 (0.08)+	0.49 (0.08)***	13
Panel B: Unweighted Average Effect Sizes									
Post-test	0.29 (0.05)***	0.29 (0.05)***	443	0.17 (0.07)*	0.17 (0.07)*	205	0.39 (0.06)***	0.39 (0.06)***	238
6 months to 1 year	0.21 (0.05)***	0.29 (0.05)***	413	0.17 (0.07)*	0.18 (0.07)*	188	0.24 (0.07)***	0.38 (0.07)***	225
> 1 year, up to 2 years	0.08 (0.11)	0.20 (0.11)+	89	0.01 (0.13)	0.08 (0.13)	59	0.22 (0.18)	0.43 (0.18)*	30
> 2 years, up to 3 years	0.08 (0.18)	0.29 (0.18)	31	0.05 (0.21)	0.09 (0.21)	22	0.16 (0.33)	0.77 (0.33)*	9
> 3 years, up to 4 years	0.16 (0.20)	0.42 (0.20)*	26	0.08 (0.35)	0.13 (0.35)	8	0.20 (0.24)	0.54 (0.24)*	18
> 4 years	0.22 (0.28)	0.51 (0.28)+	13				0.22 (0.28)	0.51 (0.28)+	13

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

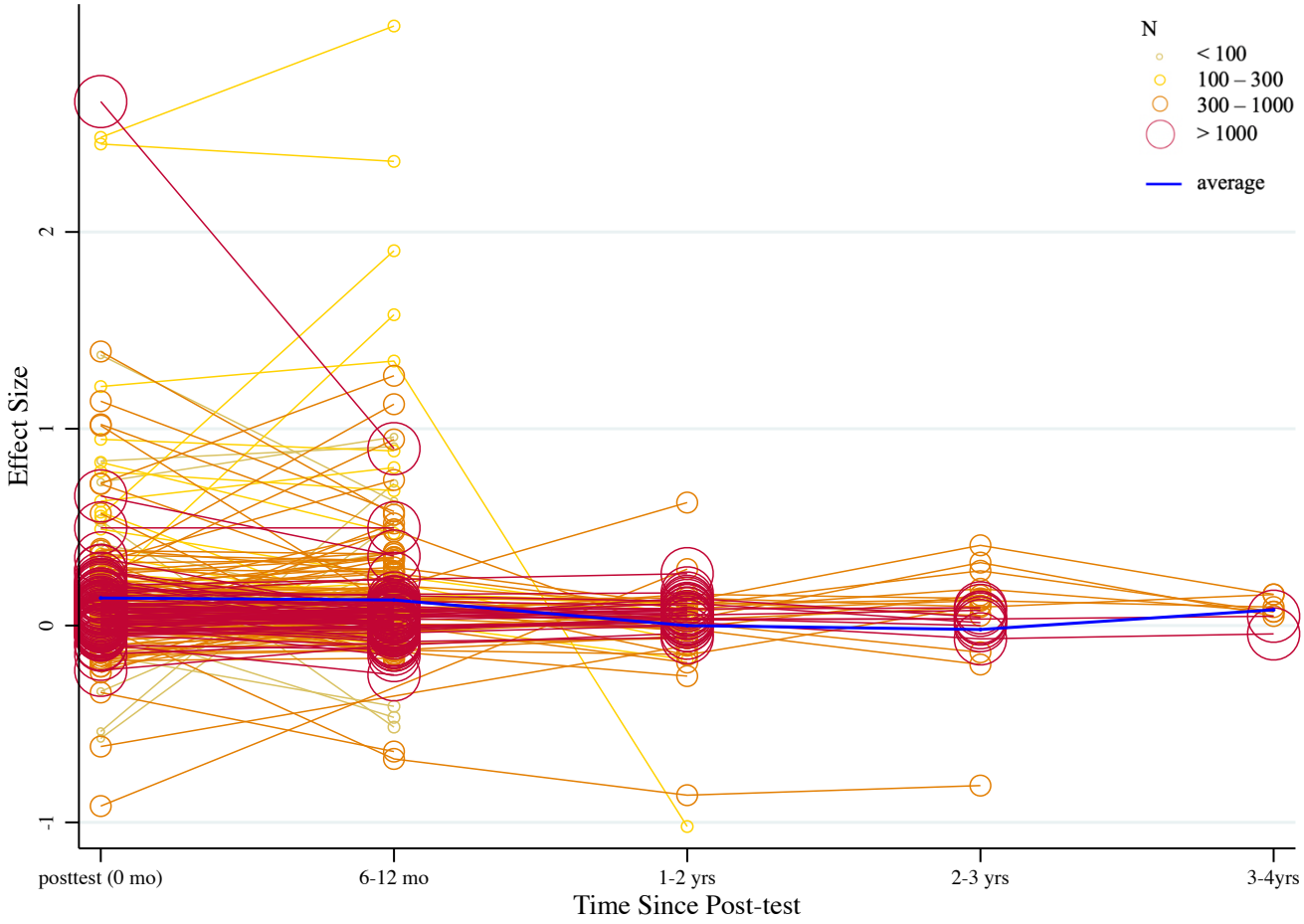
Note: "ES" = Effect size. Effect sizes are in standard deviation units. The analytic sample was constituted by "aligned groupings" which included a post-test and at least one follow-up effect size for the same construct measured using the same measure, subscales, and reporter within a treatment-control contrast. In Panel A, average effect sizes were estimated using a random effects meta-analytic model that included a random effect for study and weights ($\frac{1}{se^2}$). In Panel B, average effect sizes were estimated using a fixed effects meta-analytic model with no random or econometric fixed effect for study and no weighting. In Columns 1, 3, and 5, average effect sizes are presented for the post-test assessment wave and all follow-up assessment waves with available data. To evaluate the possibility of selection into longer-run follow-up assessments, Columns 2, 4, and 6 present average post-test effects for the outcomes collected at each follow-up wave. For example, at 6- to 12-month follow-up for all outcomes, the weighted average post-test effect size was 0.30 *sd* for the 400 outcomes collected.

Figure 3
Effect Size Trajectories- Cognitive Outcome



Note: Each line represents a cognitive construct that was measured at post-test and at least one follow-up assessment using the same measure for the same treatment-control group contrast. The average effect size trajectory is displayed in blue and was calculated with effect sizes weighted by $\frac{1}{se^2}$. As detailed in the key, coordinates were weighted by the post-test sample size (larger circles represent estimates from larger samples) with aligned color-coding (darker colors represent larger sample sizes). For display purposes, effect sizes within the -1 to 3 *SD* range are presented. As such, effect size trajectories less than -1 standard deviations at post-test are not displayed.

Figure 4
Effect Size Trajectories- Social-Emotional Outcomes



Note: Each line represents a social-emotional construct that was measured at post-test and at least one follow-up assessment using the same measure for the same treatment-control group contrast. The average effect size trajectory is displayed in blue and was calculated with effect sizes weighted by $\frac{1}{se^2}$. As detailed in the key, coordinates were weighted by the post-test sample size (larger circles represent estimates from larger samples) with aligned color-coding (darker colors represent larger sample sizes).

Table 4

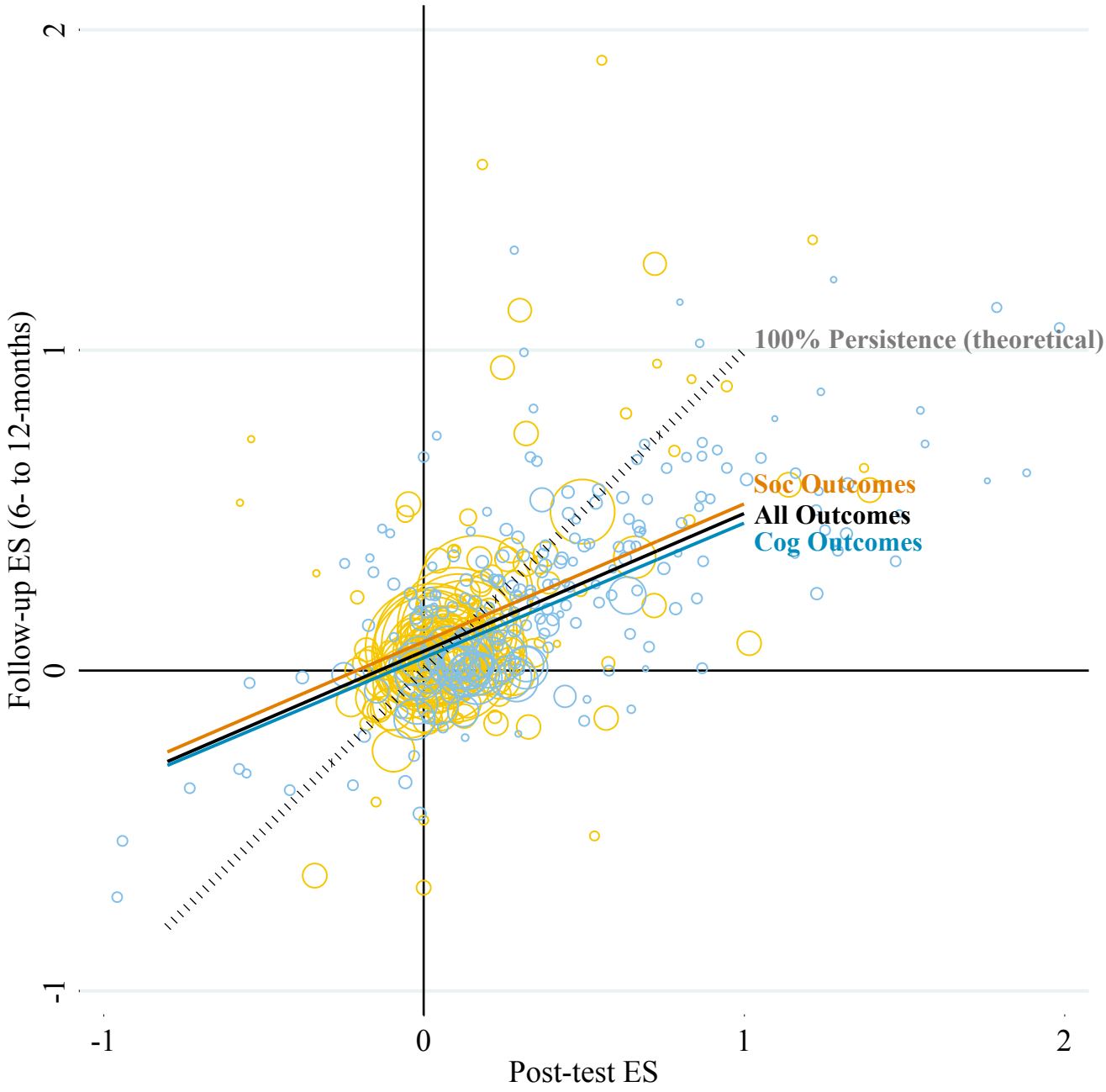
Longitudinal Persistence Rates for Social-Emotional and Cognitive Outcomes (β(se))

	All Outcomes			Split Outcomes	
	Baseline	Main Effect	Interaction	Cognitive	Social-Emotional
	(1)	(2)	(3)	(4)	(5)
Panel A: 6- to 12-months Follow-up					
Post-test Effect	0.43***	0.43***	0.42***	0.43***	0.43***
	(0.02)	(0.02)	(0.04)	(0.03)	(0.02)
Soc-Emo Outcome		0.05**	0.05**		
		(0.02)	(0.02)		
Post-test x Soc-Emo Outcome			0.01		
			(0.04)		
Constant	0.06*	0.04	0.04	0.04	0.06
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
N (study/tx contrasts/outcomes)	60 / 77 / 413	60 / 77 / 413	60 / 77 / 413	35 / 48 / 225	30 / 37 / 188
τ (study; null = .26; .19; .31)	.18	.18	.18	.12	.22
I^2 (null = 80.24%; 58.91%; 87.85%)	67.00 %	66.08 %	66.12%	19.16 %	80.03 %
Panel B: > 1 year, up to 2 years Follow-up					
Post-test Effect	-0.03	-0.03	0.37***	0.34***	-0.08+
	(0.04)	(0.04)	(0.09)	(0.09)	(0.04)
Soc-Emo Outcome		0.00	0.08*		
		(0.03)	(0.03)		
Post-test x Soc-Emo Outcome			-0.45***		
			(0.10)		
Constant	0.06	0.06	-0.05	-0.03	0.01
	(0.04)	(0.05)	(0.04)	(0.05)	(0.03)
N (study/tx contrasts/outcomes)	23 / 24 / 89	23 / 24 / 89	23 / 24 / 89	11 / 11 / 30	13 / 14 / 59
τ (study; null = .15; .17; .12)	.17	.17	.08	.06	.11
I^2 (null = 73.47%; 48.40%; 78.76%)	73.53 %	73.45 %	72.59 %	24.82 %	79.11 %

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

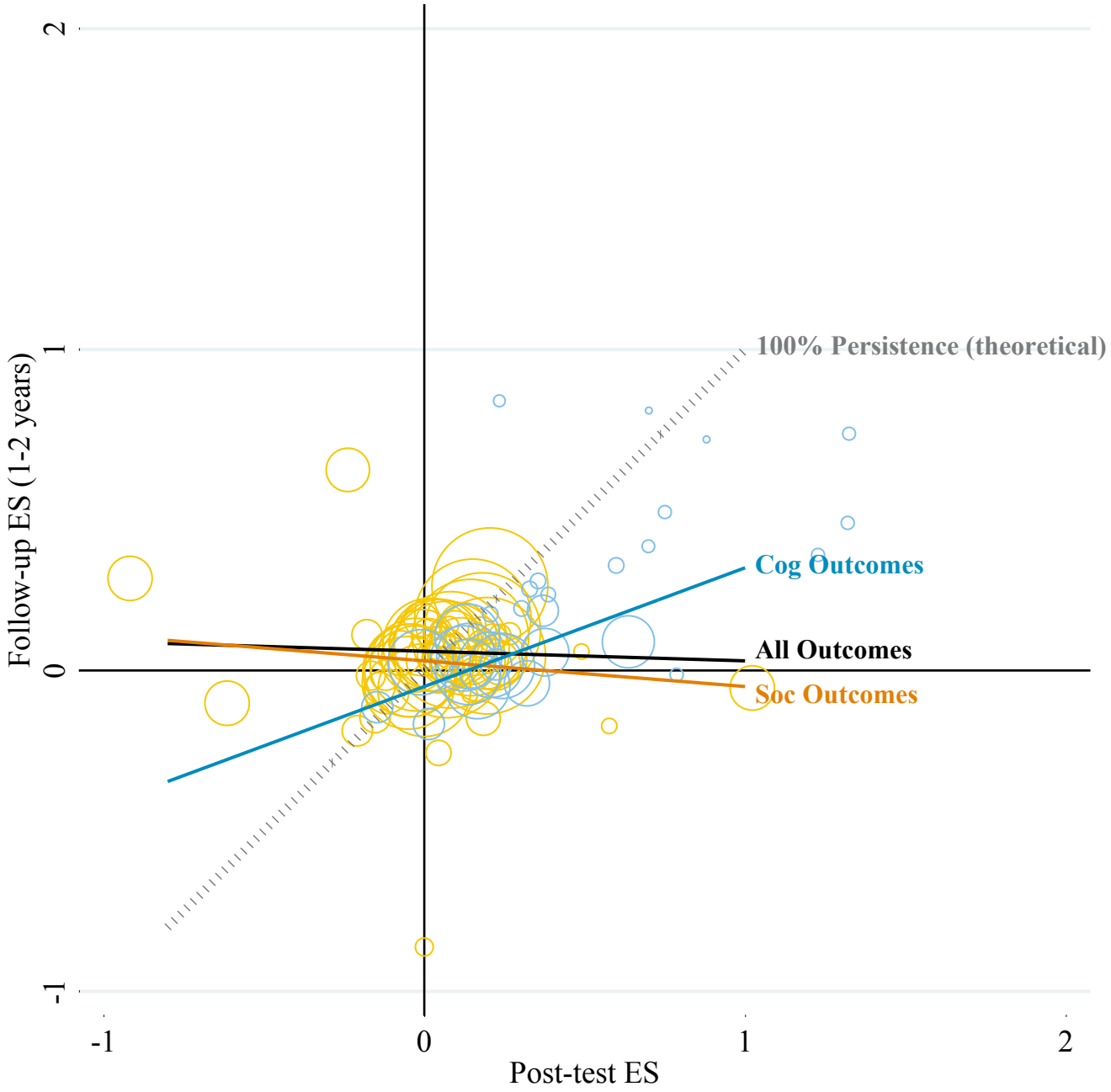
Note: "Soc-Emo Outcome" is a dummy variable for outcome type (0 = cognitive outcome, 1 = social-emotional outcome). The unit of analysis is "aligned groupings" of post-test and follow-up impacts collected for the same construct using the same measure, subscales, and reporter at post-test and at least one follow-up assessment within a study. As such, each regression tests the association between post-test and follow-up effect sizes for the same measure. Panel A presents associations between post-test and follow-up effect sizes collected 6 to 12 months after post-test. Panel B presents associations between post-test and follow-up collected greater 1 year and up to 2 years after post-test. Model parameters were estimated using a random effects meta-analytic model that included a random effect for study and weights ($\frac{1}{se^2}$). . N's are reported for studies, treatment-control group contrasts, and outcomes. Heterogeneity statistics are presented for three null models estimating follow-up effect size magnitude using random effects meta-regression. First, they're presented for the model that includes all outcomes, followed by the model including only cognitive outcomes, and finally the model including only social-emotional outcomes.

Figure 5
Persistence Patterns: Post-test to 6- to 12-month Follow-up



Note: Each bubble represents a cognitive or social-emotional construct that was measured at post-test and 6- to 12-months after post-test. Yellow bubbles represent social-emotional constructs and blue bubbles represent cognitive constructs. Model parameters were estimated using a random effects meta-analytic model that weighted effects by $\frac{1}{se^2}$ (see Table 4, Panel A). Coordinates were similarly weighted by $\frac{1}{se^2}$ (larger coordinates represent estimates with smaller standard errors at 6- to 12-months follow-up). The grey dashed "100% Persistence" line was included for reference. For display purposes, post-test and follow-up effects within the -1 to 2 *SD* range were presented. The plotted estimates are from Table 4, Panel A, Column 3.

Figure 6
Persistence Patterns: Post-test to 1- to 2-year Follow-up



Note: Each bubble represents a cognitive or social-emotional construct that was measured at post-test and 1 to 2 years after post-test. Yellow bubbles represent social-emotional constructs and blue bubbles represent cognitive constructs. Model parameters were estimated using a random effects meta-analytic model that weighted effects by $\frac{1}{se^2}$ (see Table 4, Panel B). Coordinates were similarly weighted by $\frac{1}{se^2}$ (larger coordinates represent estimates with smaller standard errors at 1- to 2-years follow-up). The grey dashed "100% Persistence" line was included for reference. For display purposes, post-test and follow-up effects within the -1 to 2 *SD* range were presented. The plotted estimates are from Table 4, Panel B, Column 3.

Table 5

Longitudinal Persistence Rates by Outcome and Intervention Types (β(se))

	Intervention Type Interactions	
	Cognitive Outcomes	Social-Emotional Outcomes
	(1)	(2)
Panel A: 6- to 12-months Follow-up		
Post-test Effect Size	0.21+	0.26
	(0.11)	(0.26)
Cog Intervention	0.06	
	(0.07)	
Soc-Emo Intervention	-0.04	0.05
	(0.11)	(0.14)
Post-test x Cog Intervention	0.26*	
	(0.12)	
Post-test x Soc-Emo Intervention	-0.34+	0.18
	(0.19)	(0.26)
Constant	0.01	0.01
	(0.06)	(0.14)
N (study/tx contrasts/outcomes)	35 / 48 / 225	30 / 37 / 188
τ (study; null = .19; .31)	.11	.23
I^2 (null = 58.91%; 87.85%)	.71 %	80.24 %
Panel B: > 1 year, up to 2 years Follow-up		
Post-test Effect Size	0.19	0.10
	(0.15)	(0.36)
Cog Intervention	-0.19	
	(0.16)	
Soc-Emo Intervention		-0.07
		(0.12)
Post-test x Cog Intervention	0.40	
	(0.26)	
Post-test x Soc-Emo Intervention		-0.18
		(0.36)
Constant	0.08	0.06
	(0.10)	(0.11)
N (study/tx contrasts/outcomes)	11 / 11 / 30	13 / 14 / 59
τ (study; null = .17; .12)	.15	.14
I^2 (null = 48.40%; 78.76%)	27.89 %	79.82 %

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

Note: "Soc-Emo Outcome" is a dummy variable (0 = cognitive outcome, 1 = social-emotional outcome). "Cog Intervention" and "Soc-Emo Intervention" are dummy variables for intervention type (reference group = "broad" interventions with both cognitive and social-emotional components). The unit of analysis is "aligned groupings" of post-test and follow-up impacts collected for the same construct using the same measure, subscales, and reporter at post-test and at least one follow-up assessment within a study. As such, each regression tests the association between post-test and follow-up effect sizes for the same measure. Panel A presents associations between post-test and follow-up effect sizes collected 6 to 12 months after post-test. Panel B presents associations between post-test and follow-up collected greater 1 year and up to 2 years after post-test. Model parameters were estimated using a random effects meta-analytic model that included a random effect for study and weights ($\frac{1}{se^2}$). N's are reported for studies, treatment-control group contrasts, and outcomes. Heterogeneity statistics are presented for two null models estimating follow-up effect size magnitude using random effects meta-regression. First, they're presented for the model including only cognitive outcomes, and second for the model including only social-emotional outcomes.

Supplemental File for:

"Do Intervention Impacts on Social-Emotional Skills Persist at Higher Rates than Impacts on Cognitive Skills?
A Meta-Analysis of Educational RCTs with Long-Term Follow-up"

Additional Methodological Details

A full meta-analysis protocol is available on openICPSR ([LINK TO BE INSERTED PRIOR TO PUBLISING]) that further details specific protocols employed in the study inclusion/exclusion, coding, data cleaning, and effect size calculation processes.

Inclusion Determinations

Note that there were 10 duplicate papers included in the original sample of papers from the eight meta-analyses (i.e., there were 436 unique papers pre-removal).

RCT with Behavioral or Cognitive Outcomes

The first author and at least one Masters-level student independently reviewed each intervention to determine whether each was an RCT (first) and then, of these, whether each RCT reported behavioral or cognitive outcomes.[7] Discrepancies in determinations were resolved by the first author and a project PI.

Follow-up at Least 6 months after Post-test on the Same Sample

First, for each intervention, a Ph.D. student determined the paper that reported "initial impacts" (i.e., the first paper reporting intervention impacts). Second, at least two research assistants used these initial impact papers to conduct a Google Scholar search to identify and gather all additional papers reporting treatment impacts for that intervention. Next, at least two research assistants independently reviewed all of the gathered papers for each intervention and determined whether the intervention contained adequate follow-up. The first author reviewed all decisions and resolved discrepancies.

Usable Data

The first author reviewed all of the papers for the availability of usable data. A Masters-level student reviewed cases the first author deemed to require exclusion.

Coding

The coding protocol can be found in the meta-analysis protocol on openICPSR ([LINK TO BE INSERTED PRIOR TO PUBLISING]). The protocol details guiding principles of the coding process as well as code-level details that informed how the coders made determinations about each code.

Construct Categorizations

Following coding, the master coder and a study PI independently conceptually categorized the author-reported constructs for each reported treatment impact. The team reviewed the constructs and derived categories that conceptually captured the key constructs present in the data. The master coder and study PI then independently categorized each construct according to the following options: achievement composite, attendance, criminality, educational attainment, externalizing behaviors, general cognition, grades/GPA, internalizing symptoms, language and literacy, learning skills, math, mixed composite (i.e., a measure that combined cognitive and social-emotional skills), other academic ability, retention, general social-emotional skills, special education designation, and substance use. Some outcomes did not fall within these categories and were set to be missing a construct categorization.

Intervention Target Categorizations

Intervention targets were coded based on skills that the study authors explicitly stated as skills that the intervention aimed to improve (see coding protocol for details). To be categorized

[7] In several cases there was more than one paper that reported treatment impacts on the same intervention. In these cases, all papers on a particular intervention were reviewed to determine intervention inclusion or exclusion.

as a cognitive intervention for the purposes of the exploratory outcome by intervention type interaction analyses, an intervention had to target no social-emotional skills and at least one of the following cognitive skills: math, language and literacy, executive function, general cognition, or science. Alternately, to be categorized as a social-emotional intervention the intervention had to target no cognitive skills and at least one of the following social-emotional skills: social-emotional skills, learning skills, substance use, or psychological well-being. To be categorized as an intervention with broad targets, the intervention had to target at least one cognitive skill and one social-emotional skill.

Effect Size and Standard Error Calculations

The first author worked closely with the study PIs in determining effect sizes. In keeping with our "double coding" process, an additional Ph.D. student checked all calculations. Figure S1 details the formulas used to calculate effect sizes based on the available, reported results.

The ultimate goal of the effect size calculation process was to identify one effect size for each coded outcome. While the standard protocol was to calculate effect sizes according to the formula detailed in the manuscript, or to use a viable author-reported effect sizes when these were available, there were many cases in which additional decision criteria were used to determine which effect size to use, or to calculate the effect size.

Adjustments for Effect Sizes Calculated using SEs, t-statistics, and f-statistics

In cases when standard deviations were not provided and viable reported effect sizes were not available, reported standard errors, *t*-statistics, and *f*-statistics were used to derive effect sizes (see Figure S1). In the case that any of these statistics were used to calculate effect sizes for a given outcome, the first author returned to the original paper to check whether the statistic appeared to have been calculated in a model with the inclusion of the pre-test control. In these cases, an adjustment was made when calculating the effect size given the likelihood that standard errors may have be artificially reduced as a result of the inclusion of this control, thus biasing the effect sizes calculated using these estimates. In the cases that this control was included, the standard errors calculated from the available statistics were divided by the square root of 1 minus R^2 (assuming an R^2 between pre- and post-test measures of .50) in the effect size calculation process (using the formulas outlined in Figure S1). Thus, adjustments were made by dividing standard errors by .87 in these cases to ensure that the standard errors were not inaccurately small in the effect size calculation process.

Importantly, in many cases these adjusted effect sizes were then used to estimate an accompanying standard error for use in our models (i.e., to up weight studies with greater precision). To ensure that these estimated standard errors used were not inaccurately large in our meta-analytic models due to the .87 effect size adjustment, estimated standard errors were multiplied by .87.

Calculating ES using P-values

In the case that no alternative statistics were available to use in calculating effect sizes, the last resort was to estimate an effect size using reported *p*-values. If precise *p*-values were reported (e.g., ".002"), then *t*-statistics were calculated from these *p*-values and the formulas detailed in Figure S1 were then used to convert *t*-values to effect sizes.

If relatively precise *p*-values were reported (e.g., "< .05"), we found the smallest difference between means for each measure within a given study and assumed this *p*-value was the largest possible associated *p*-value (e.g., .05). For these cases, we then converted the *p*-value to a *t*-value using the "invt" function in Stata, assuming a two-tailed test (i.e., we divided the *p*-

value by 2). Next, we calculated the effect size from this *t*-value (as described above), and recovered a SD from this calculated effect size. For the cases in which the same measure was available within a study but did not qualify as having the smallest difference between means, the recovered SD was then used to calculate these effect sizes.

In the case that *p*-values were only reported to be statistically non-significant, with no precise value associated, we found the largest difference between means for each measure within a given study and assumed that this *p*-value was .10. We then converted the *p*-value to a *t*-value using the same procedure described above for relatively precise *p*-values and recovered a SD that was then used to calculate the effect size for the other cases within a study that had smaller differences between the means for each measure.

In the cases where treatment and control group means were not provided for an outcome, and the treatment impact was noted to be statistically non-significant, *p*-values were assumed to be .10 and *t*-statistics were calculated from these *p*-values. Because means were not available, an alternate equation was used to convert *t*-values to effect sizes (see next section).

For all of these aforementioned processes, we made the .87 pre-test covariate adjustment when it appeared that the *p*-value came from a model including a pre-test control (see previous section for more details).

Calculating ES from F- and T-statistics when Means were Not Reported.

When treatment and control group means were not provided, and effect sizes were estimated using *t*-statistics (only in the case of *p*-value conversions) or *f*-statistics (in the case of one study), the following equations were used:

$$ES = t \times \sqrt{\frac{n_{tx} + n_{cntrl}}{n_{tx} * n_{cntrl}}}$$

$$ES = \sqrt{f} \times \sqrt{\frac{n_{tx} + n_{cntrl}}{n_{tx} * n_{cntrl}}}$$

Choosing between Using Author-Reported or Calculated Effect Sizes

In cases when both author-reported effect sizes and calculated effect sizes were available for an outcome, we opted for consistency in using either reported or calculated effect sizes for all outcomes in a paper, if possible. For example, if a particular paper reported means and standard deviations for 20 outcomes that we used to calculate effect sizes, and also reported viable effect size estimates for 10 of those outcomes, we opted to use our calculated effect sizes for all outcomes because these were available consistently.

In cases when within-paper consistency was not an issue, we then checked for differences in reported effect sizes and calculated effect sizes. If the difference in estimates was less than 1 *SE* for all effect sizes within a paper, we opted to use the reported effect size because this estimate was, presumably, more precise if authors incorporated controls for baseline covariates or other relevant covariates in their estimations. If the difference in estimates was greater than 1 SE for any outcome within a paper, the first author checked whether issues related to valence (see next section) may have driven differences in the final reported and calculated effects. The first author also determined whether there were any issues (e.g., longitudinal effects were modeled linearly in a growth curve model, interaction terms were included in the model, etc.) in the estimation strategies used to calculate the author-reported effect size that the coders missed in

the coding process (i.e., only "viable" effect sizes should have been coded). The first author reviewed decisions with one of the study PI to arrive at final determinations about whether to use the reported or calculated effect sizes. So long as there were no estimation issues with the reported effect sizes, these were used with the assumption that such effects should be more precise due to the inclusion of covariates when modeling the estimates.

Determining Effect Size Valence

Unfortunately, we failed to code for effect size valence in the primary coding process. Thus, a post-coding process was initiated to identify the valence of each effect size included in the meta-analysis. For each effect size, the first author and a study PI independently determined whether each effect size should be multiplied by 1 or -1 to capture that a higher score on the construct was positive (e.g., math scores) or negative (e.g., depressive symptoms), respectively. With the addition of the second study PI, the team reviewed all discrepant cases and the team resolved discrepancies. For the effect sizes that the team couldn't reach resolution on, two research assistants, at least at the Masters-student level, reviewed each case by returning to the respective paper and gathering evidence for a valence determination. The first author reviewed these cases and made final determinations. The study PI were consulted for complicated cases.

Calculated effect sizes were multiplied by the valence. In the case that a reported effect size was used in analyses, however, an additional round of valence coding was initiated to identify whether the reported effect sizes were already re-valanced (e.g., if a study found reduced behavioral problems (lower mean), they reported a positive treatment impact), or whether effect sizes were presented as expected given their measure valence (i.e., if a study found reduced behavioral problems (lower mean), they reported a negative treatment impact). Three Masters-level research assistants reviewed all of the reported effect sizes that were suspected to have a high likelihood of valence-related issues (e.g., social-emotional outcomes). Two Masters-level research assistants reviewed all of the reported effect sizes that were not likely to have valence-related issues (e.g., cognitive outcomes). The first author reviewed these cases and made a final determination.

Negative Post-test Effect Sizes

As an additional check, the team reviewed the valence of outcomes for which the post-test effect size was negative and statistically significant after valence adjustments were made. Given the unlikelihood that treatments produce a negative, statistically significant post-test impact, we hoped that this check would catch errors in valence coding. There were 57 cases of statistically significant, negative post-test effects. Either two Masters-level research assistants and one Ph.D.-student-level research assistant, or the first-author and one Ph.D.-student-level research assistant, reviewed these cases. For each case, the reviewers indicated when valence should be re-coded and effect sizes should be adjusted, accordingly. The paper's first author reviewed their determinations and resolved discrepancies as needed. 7 cases were identified as needing valence re-coding and were re-coded.

Results Presented for Subsamples & Multiple Treatment Groups

Notably, there were cases when data were reported separately for different sub-samples within a study (e.g., for boys and girls, for "low-risk" and "high-risk" participants, etc.). For these cases, we derived a main treatment effect by averaging the effect size estimates for each group, weighted by the group sample size. The same weighted averaging was used for standard errors and *p*-values. Critically, if the treatment effect was only reported for one sub-sample (e.g., only boys, only "low-risk" participants, etc.), then the effects were dropped from the meta-

analysis so that each estimate represented a main treatment impact of original random assignment to treatment or control.

Results were also commonly reported for multiple treatment groups formed via random assignment within a study. We opted to leave effect sizes presented separately by treatment group when possible since the effects reflected experimental treatment impacts. However, there were some instances when effect sizes were reported for each treatment group separately at earlier assessment waves (e.g., pre-test, post-test, 6-12-month follow-up), and in aggregated form at later assessment waves (e.g., 3-year follow-up). In these cases, treatment-specific effect sizes, standard errors, and *p*-values were averaged to form an average treatment effect that could be investigated in alignment with the effect sizes from later assessment waves.

Additional Analytic Details

Sensitivity Analyses

Below we detail additional analytic details for two sensitivity analyses presented in the manuscript. The first pertains to an analysis that adjusted standard errors to account for cluster-based randomization, and the second is for an analysis that removed effect sizes calculated using a heavy reliance on estimation.

Cluster-Related Standard Error Adjustment

We performed an analysis in which we adjusted the standard errors to account for clustering concerns. To do so, we multiplied the standard errors by the square root of a variance inflation factor, acknowledging that in some cases such an adjustment *would* be appropriate (i.e., when cluster adjustments were not yet made), and in other cases the adjustment would be too conservative (i.e., when cluster adjustments were already made and/or when pre-test covariates were accounted for in calculating treatment impacts). The VIF was calculated as follows:

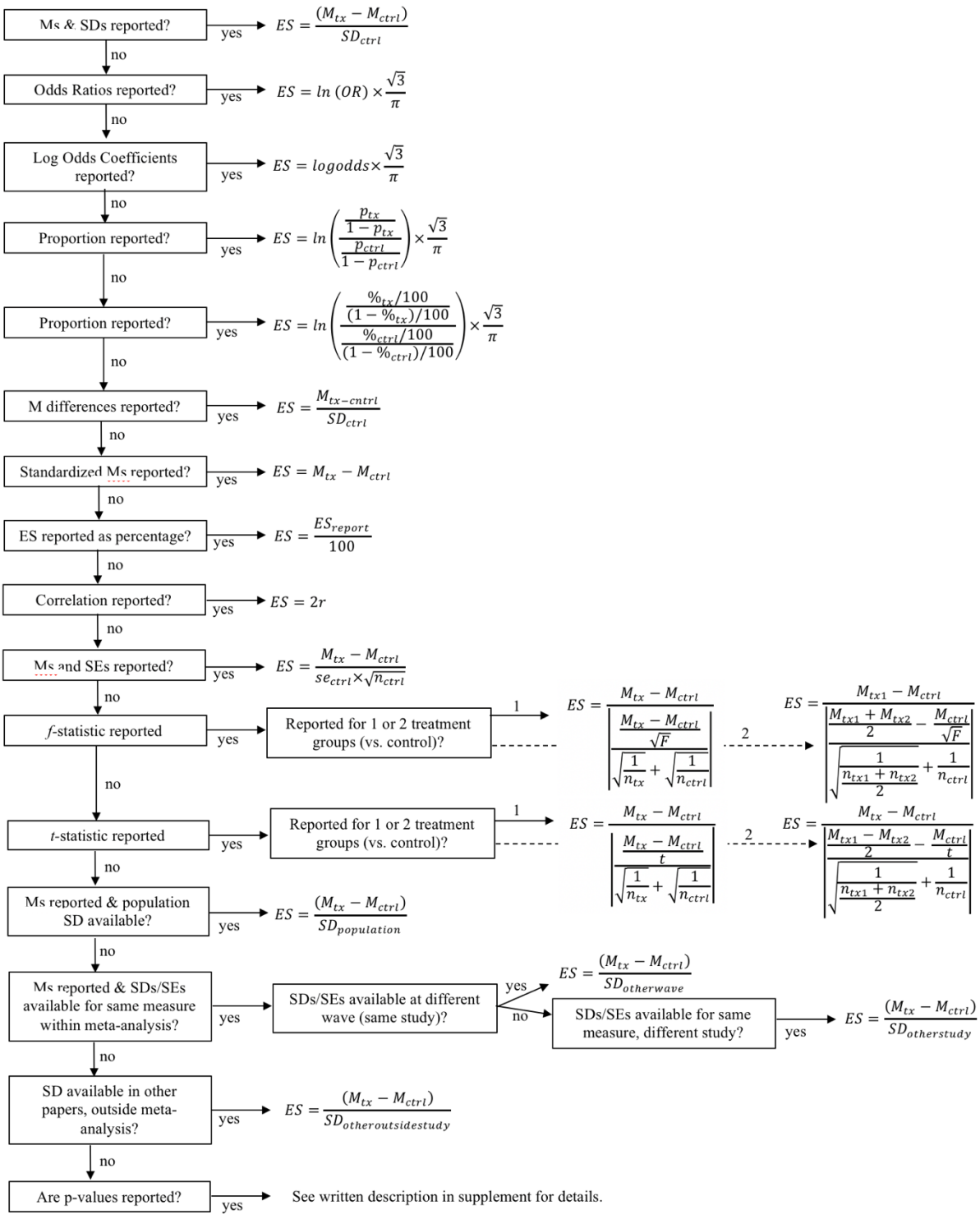
$$VIF = 1+ (ICC + m -1)$$

where ICC = 0.10 and m represented the number of participants within a cluster. Because we did not code for how many clusters were randomly assigned, we tested two models: one in which we assumed that there were 20 clusters and one in which we assumed there were 40 clusters. Thus, to calculate m, we divided the baseline sample size by 20 and 40.

Removing "Estimated" Effects

We performed an analysis in which we removed effect sizes that required more-than-typical estimation in the calculation process. First, we dropped effect sizes calculated for dichotomous outcomes because the conversion from odds to standard deviation units depends on the distribution underlying the dichotomous outcome, which may violate normality in some cases. This included effect sizes calculated through transforming raw data reported as percentages and proportions as well as through transforming effect sizes reported as Odds Ratios and Log Odds Coefficients. Second, we dropped effect sizes that were calculated from imprecise p-values (see "Calculating ES using P-values"). Finally, we dropped cases in which standard deviations were estimated from: a) population level SDs, 2) SDs reported for the same measure within our meta-analytic, or 3) SDs reported in other papers outside of the meta-analytic sample. Effect sizes were dropped prior to forming the aligned analytic groupings for use in analyses.

Figure S1
Effect Size Calculation Flow Chart



Note: This flow chart details the formulas used for calculating effect sizes and the decision-making process for deciding which calculation to use. Additional details relevant to this process, including adjustments to standard errors when these were estimated in models controlling for pre-test scores, and the procedure used to calculate effect sizes from p-values (if no other information was provided) is included in the supplemental text.

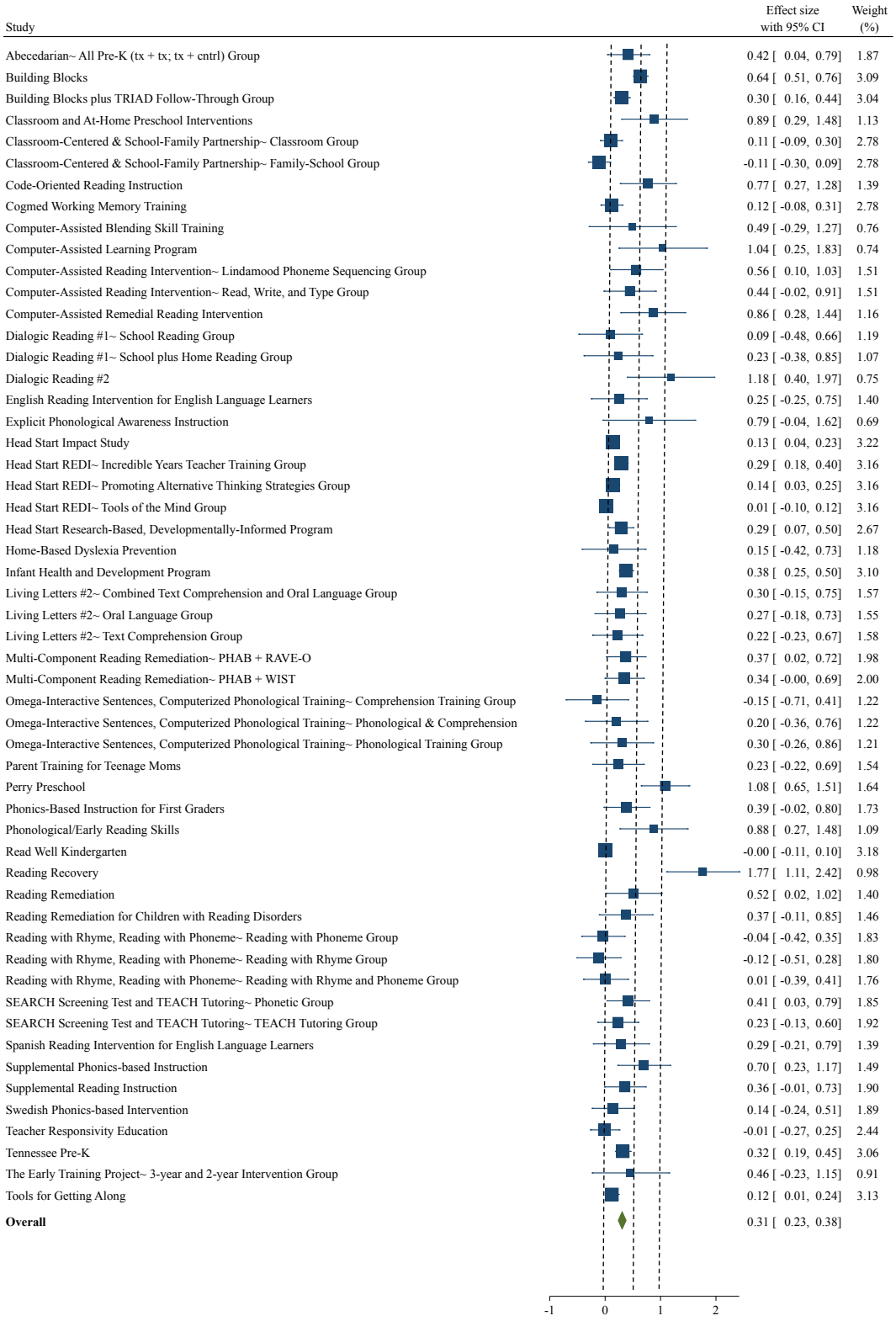
Table S1
Construct Categorization Examples

	Example Construct(s)	Example Measure(s)
Cognitive		
Achievement Composite	Pre-Academic Achievement	Composite of Woodcock Johnson subscales (letter-word identification, spelling, and applied problems)
General Cognition	IQ	Stanford-Binet Intelligence Scale
	Verbal Short-term Memory	Automated Working Memory Assessment
Language and Literacy	Vocabulary	Peabody Picture Vocabulary Test
	Auditory-Vocal Association	Illinois Test of Psycholinguistic Abilities
Math	Arithmetic	Wechsler
	Calculation	Woodcock Johnson
Other Academic Abilities	Science	Tennessee Comprehensive Assessment Program
	Social Studies	Stanford Achievement Tests
Social-Emotional		
Crime	Lifetime Violent Arrests	Study-created measures
	Convictions	
Externalizing Behaviors	Aggressive Behaviors	Child Behavior Checklist
	Disruptive Behavior	Finn Disruptive Behavior Scale
Internalizing Symptoms	Anxiety	Children's Manifest Anxiety Scale
	Depression	Child Depression Inventory
General Social-Emotional Skills	Social Skills	Social Skills Rating Scale
	Attributional Style	Children's Attributional Style Questionnaire
Substance Use	Alcohol Consumption	Study-created measure
	Anti-Marijuana-Use Attitudes	Teenager's Self Test

Note: Examples of the constructs and measures that were categorized in each of the construct categories.

Figure S2

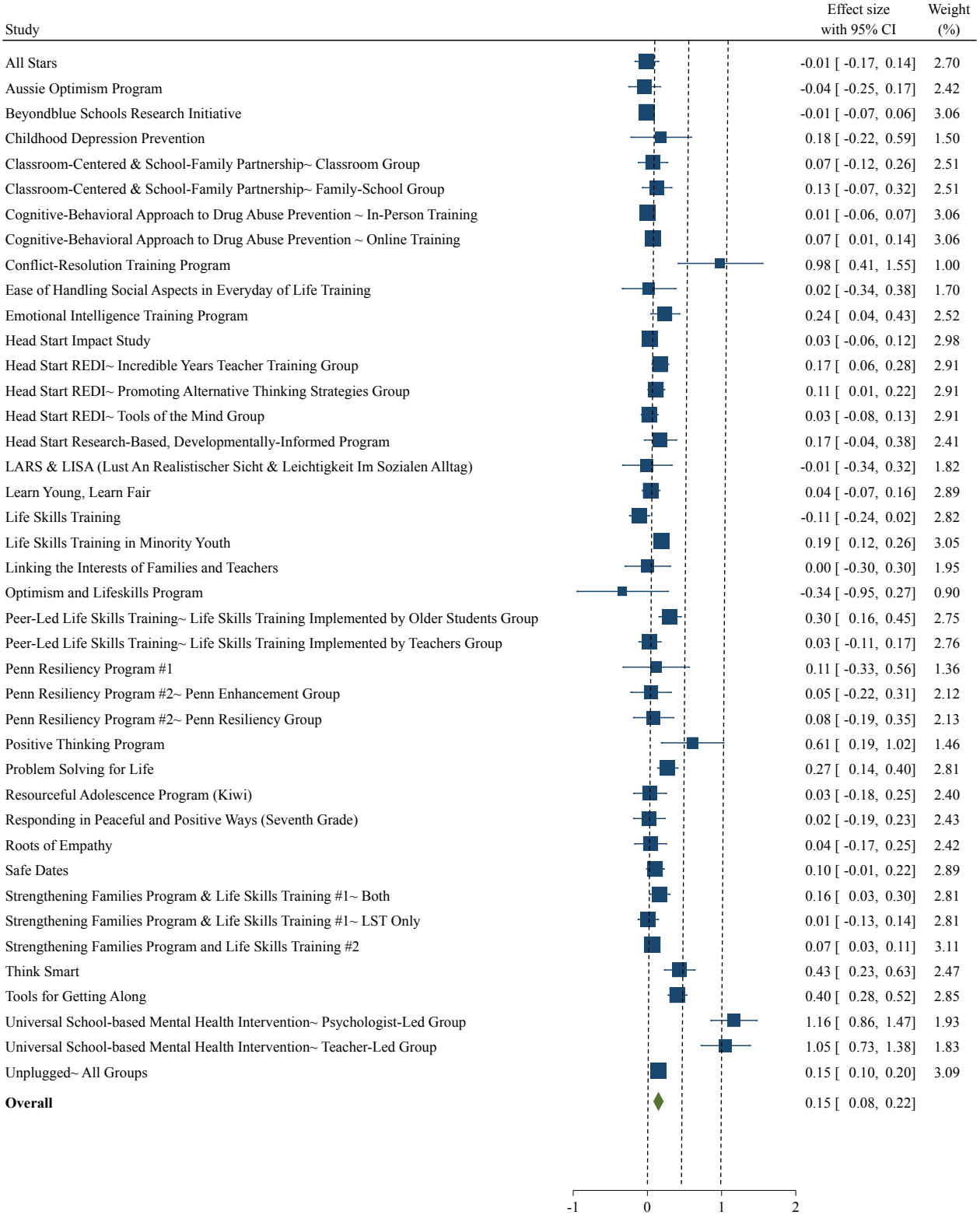
Forest Plot: Average Cognitive Outcomes at Post-test (unweighted)



Note: Unweighted average post-test effect sizes for all cognitive outcomes included in the analytic sample, averaged within each experimental group.

Figure S3

Forest Plot: Average Social-Emotional Outcomes at Post-test (unweighted)



Note: Unweighted average post-test effect sizes for all social-emotional outcomes included in the analytic sample, averaged within each experimental group.

Table S2

Longitudinal Persistence Rates by Outcome for Long-Term Follow-up Greater than 2 Years after Post-test (β(se))

	RE, weighted (1)
> 2 years Follow-up	
Post-test Effect	0.16
	(0.15)
Soc-Emo Outcome	-0.10
	(0.12)
Post-test x Soc-Emo Outcome	0.30
	(0.32)
Constant	0.07
	(0.09)
N (study/tx contrasts/outcomes)	18 / 21 / 53
τ (study; null model= .15)	.17
I^2 (null model= 33.09%)	34.86%

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

Note: "Soc-Emo Outcome" is a dummy variable for outcome type (0 = cognitive outcome, 1 = social-emotional outcome). The unit of analysis is "aligned groupings" of post-test and follow-up impacts collected for the same construct using the same measure, subscales, and reporter at post-test and at least one follow-up assessment within a study. As such, each regression tests the association between post-test and follow-up effect sizes for the same measure. Model parameters were estimated using a random effects meta-analytic model that included a random effect for study and weights ($\frac{1}{se^2}$). N's are reported for studies, treatment-control group contrasts, and outcomes.

Table S3

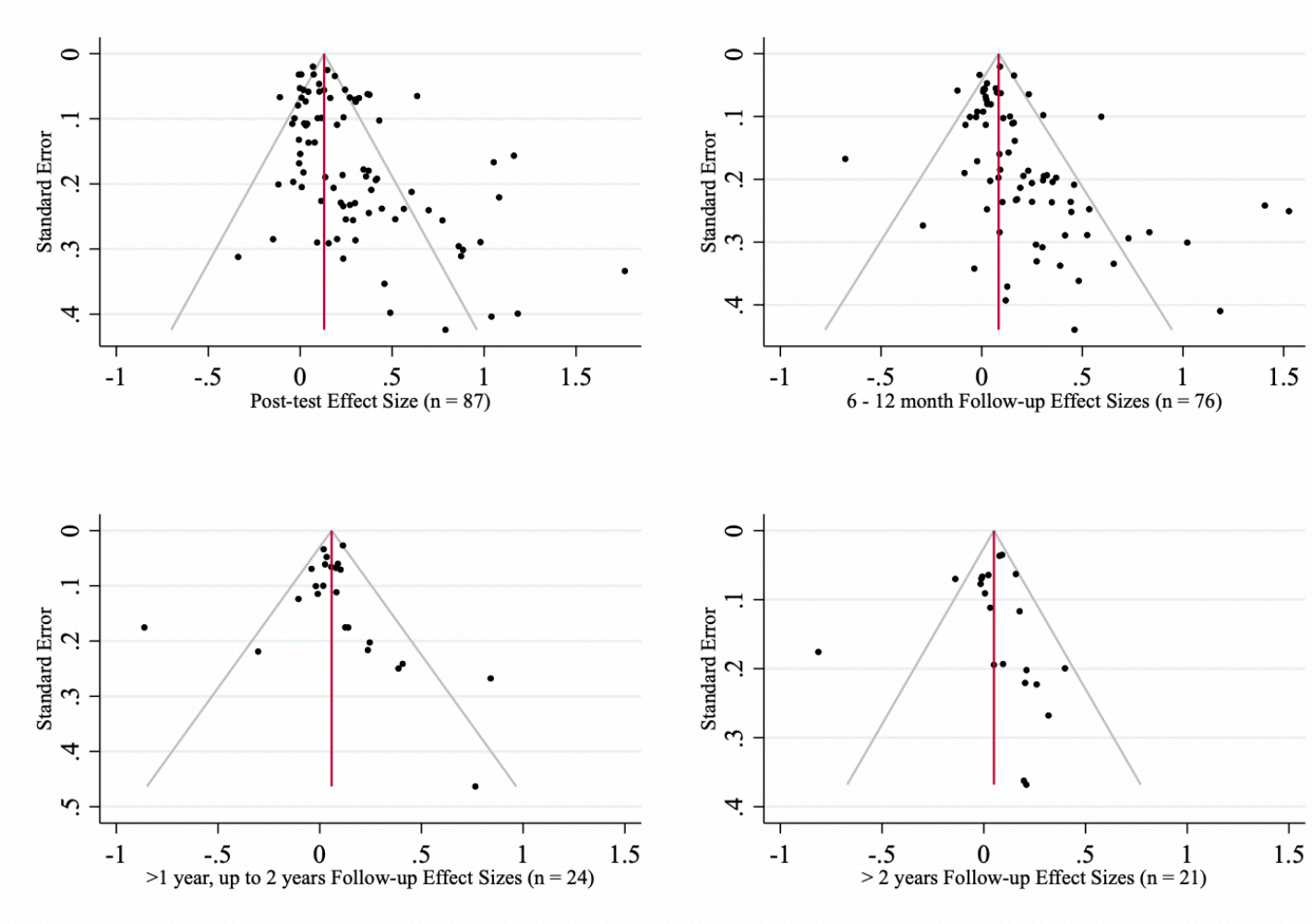
Robustness Checks for Longitudinal Persistence Rates by Outcome (β(se))

	Fixed Effects (1)	Age Covariate (2)	Intensity (< 200 hours) (3)	Intensity (< 100 hours) (4)	Unweighted (5)	Cluster Covariate (6)	Cluster Weighting (7)	Negative Post-tests Dropped (8)	Estimated Effects Dropped (9)
Panel A: 6- to 12-month Follow-up									
Post-test Effect	0.42***	0.42***	0.43***	0.45***	0.42***	0.42***	0.46***	0.42***	0.43***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.12)	(0.04)	(0.04)	(0.05)	(0.04)
Soc-Emo Outcome	0.06**	0.05**	0.05**	0.07***	-0.04	0.05**	0.06*	0.06*	0.05*
	(0.02)	(0.02)	(0.02)	(0.02)	(0.11)	(0.02)	(0.02)	(0.02)	(0.02)
Post-test x Soc-Emo Outcome	0.01	0.02	0.01	-0.01	0.27	0.02	-0.01	-0.01	0.03
	(0.04)	(0.04)	(0.04)	(0.05)	(0.22)	(0.04)	(0.06)	(0.06)	(0.04)
Age		0.00							
		(0.00)							
Cluster						-0.06			
						(0.06)			
Constant	--	0.12+	0.04	0.02	0.08	0.07+	0.03	0.04	0.04
		(0.07)	(0.03)	(0.03)	(0.08)	(0.04)	(0.03)	(0.03)	(0.03)
N (study/tx contrasts/outcomes)	60 / 77 / 413	59 / 76 / 407	56 / 73 / 404	52 / 67 / 352	60 / 77 / 413	60 / 77 / 413	60 / 77 / 413	59 / 75 / 329	56 / 72 / 365
Panel B: 1- to 2-year Follow-up									
Post-test Effect	0.29+	0.36**	0.33**	0.33**	0.46	0.36***	0.37***	0.46***	0.30***
	(0.16)	(0.11)	(0.12)	(0.12)	(0.48)	(0.11)	(0.08)	(0.12)	(0.08)
Soc-Emo Outcome	0.07*	0.08*	0.07+	0.07+	0.02	0.08*	0.08**	0.08+	0.05*
	(0.04)	(0.03)	(0.03)	(0.03)	(0.31)	(0.03)	(0.03)	(0.04)	(0.02)
Post-test x Soc-Emo Outcome	-0.37*	-0.43***	-0.42**	-0.42**	-0.75	-0.44***	-0.52***	-0.41**	-0.09
	(0.16)	(0.12)	(0.13)	(0.13)	(0.66)	(0.11)	(0.12)	(0.15)	(0.13)
Age		0.00							
		(0.00)							
Cluster						-0.07			
						(0.06)			
Constant	--	0.00	-0.03	-0.03	0.01	-0.01	-0.03	-0.08	-0.03
		(0.09)	(0.04)	(0.04)	(0.28)	(0.06)	(0.02)	(0.05)	(0.02)
N (study/tx contrasts/outcomes)	23 / 24 / 89	22 / 23 / 83	21 / 22 / 83	21 / 22 / 83	23 / 24 / 89	23 / 24 / 89	23 / 24 / 89	23 / 24 / 69	19 / 20 / 69

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

Note: The unit of analysis is "aligned groupings" of post-test and follow-up impacts collected for the same construct using the same measure, subscales, and reporter at post-test and at least one follow-up assessment within a study. Column 1 presents a fixed effects meta-analytic model with an econometric fixed effect for study and weights ($\frac{1}{se^2}$). Column 2 presents the primary random effects meta-analytic model with a covariate for participant age at pre-test. Columns 3 and 4 present the primary model with analytic samples limited to interventions with less than 200, or less and 100 hours of intervention. Column 5 presents a fixed effects meta-analytic model with no random or econometric fixed effect and no weights. Column 6 presents the primary model with a covariate for whether cluster randomization was used. Column 7 presents the primary model with a standard error adjustment for clustering. Column 8 presents the primary model with negative post-test effects dropped. Column 9 presents the primary model with effect sizes that relied on estimation in the calculation process dropped.

Figure S4
Funnel Plots



Note: Grey lines represent 95% confidence intervals. Each coordinate represents the average effect size for each treatment-control group contrast contributing aligned constructs to the analytic sample (for which the same construct was measured using the same measure at post-test and at least one follow-up assessment).

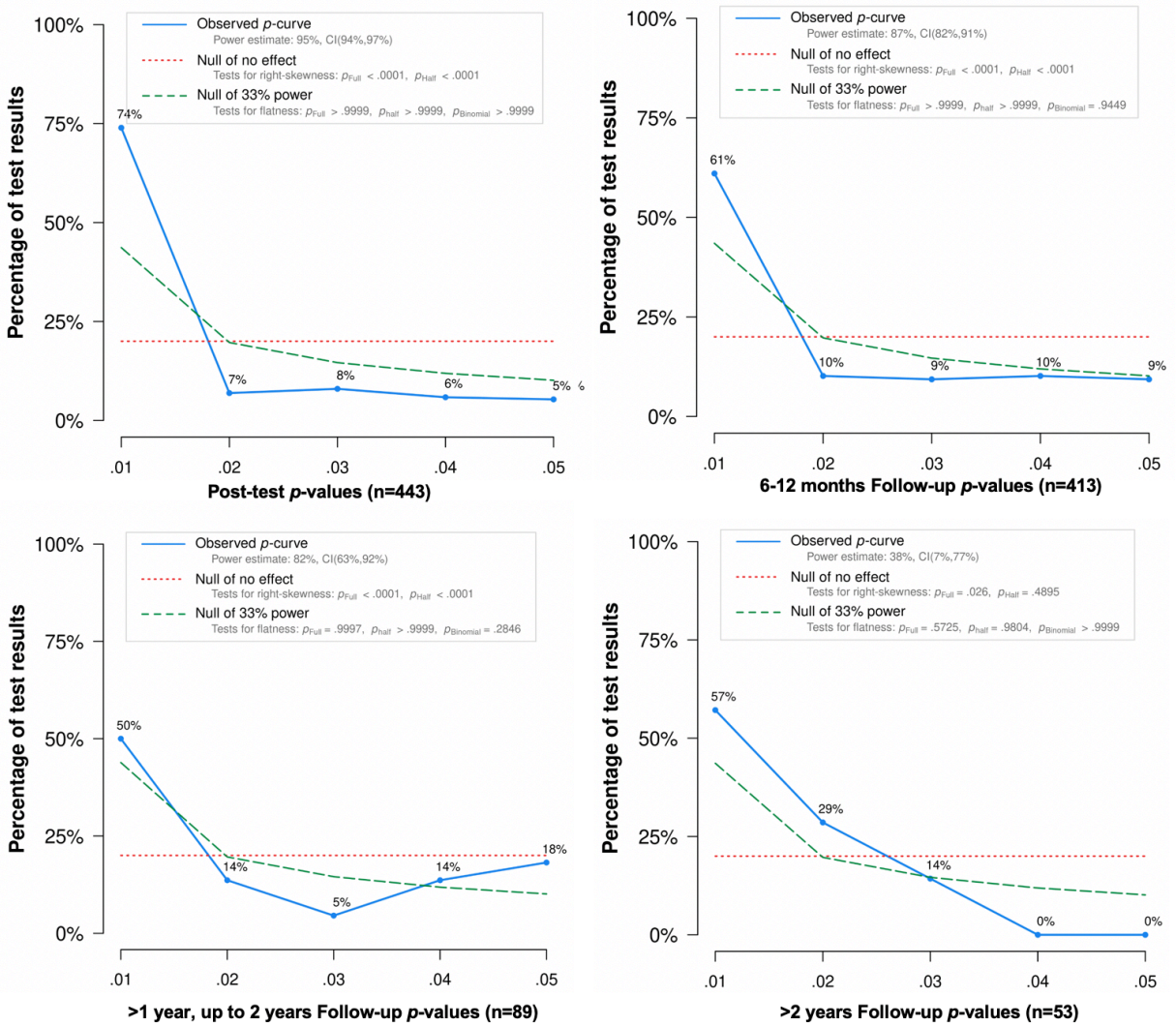
Table S4
PEESE Test (β(se))

	PEESE Test (1)
Panel A: 6- to 12-months Follow-up	
Post-test Effect	0.39***
	(0.04)
Soc-Emo Outcome	0.06**
	(0.02)
Post-test x Soc-Emo Outcome	0.04
	(0.04)
Standard Error	1.08***
	(0.26)
Constant	-0.13**
	(0.05)
N (study/tx contrasts/outcomes)	60 / 77 / 413
Panel B: > 1 year, up to 2 years Follow-up	
Post-test Effect	0.35**
	(0.11)
Soc-Emo Outcome	0.07*
	(0.03)
Post-test x Soc-Emo Outcome	-0.44***
	(0.11)
Standard Error	0.35
	(0.42)
Constant	-0.08
	(0.05)
N (study/tx contrasts/outcomes)	23 / 24 / 89

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

Note: This table presents alternate grouping approaches. In (1), the analytic sample was constituted of the same "aligned groups", but measure, subscale, and reporter were allowed to vary across waves. In (2), the analytic sample was constituted of all social-emotional and all cognitive outcomes averaged together at each wave (e.g., charting the average of all social-emotional impacts at post-test to the average of all social-emotional impacts at follow-up). "Soc Outcome" is a dummy variable for outcome type (0 = cognitive outcome, 1= social-emotional outcome). Parameters were estimated using a random effects meta-analytic model that included a random effect for study and weights ($\frac{1}{se^2}$).

Figure S5
Relative Frequency of P-values < .05



Note: Each figure contains all of the *p*-values included in the analytic sample at post-test and each follow-up wave. Post-tests and follow-up impacts are from the larger analytic sample which was constituted of "aligned groupings" in which the same construct was measured using the same measure, subscales, and reporter at post-test and at least one follow-up assessment within a study. *P*-curve figures were created on p-curve.com (Simonsohn, Nelson, & Simmons, 2015).

Table S5

Longitudinal Persistence Rates by Outcome using Alternate Grouping Approaches (β(se))

	Alternate Grouping #1 (1)	Alternate Grouping #2 (2)
Panel A: 6- to 12-months Follow-up		
Post-test Effect	0.42***	0.45***
	(0.03)	(0.09)
Soc-Emo Outcome	0.04*	-0.04
	(0.02)	(0.03)
Post-test x Soc-Emo Outcome	0.02	0.40**
	(0.04)	(0.14)
Constant	0.05	0.03
	(0.03)	(0.03)
N (study/tx contrasts/outcomes)	62 / 80 / 411	69 / 90 / 98
Panel B: > 1 year, up to 2 years Follow-up		
Post-test Effect	0.35***	0.39*
	(0.10)	(0.19)
Soc-Emo Outcome	0.07*	0.04
	(0.03)	(0.07)
Post-test x Soc-Emo Outcome	-0.43***	-0.54
	(0.10)	(0.38)
Cluster		
Constant	-0.04	0.00
	(0.04)	(0.07)
N (study/tx contrasts/outcomes)	24 / 25 / 90	29 / 32 / 33

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

Note: This table presents alternate grouping approaches. In (1), the analytic sample was constituted of the same "aligned groups", but measure, subscale, and reporter were allowed to vary across waves. In (2), the analytic sample was constituted of all social-emotional and all cognitive outcomes averaged together at each wave (e.g., charting the average of all social-emotional impacts at post-test to the average of all social-emotional impacts at follow-up). "Soc Outcome" is a dummy variable for outcome type (0 = cognitive outcome, 1= social-emotional outcome). Parameters were estimated using a random effects meta-analytic model that included a random effect for study and weights ($\frac{1}{se^2}$).

References for Papers Included in MERF

Learn Young, Learn Fair

Kraag, G., Van Breukelen, G. J. P., Kok, G., & Hosman, C. (2009). 'Learn Young, Learn Fair', a stress management program for fifth and sixth graders: Longitudinal results from an experimental study. *Journal of Child Psychology and Psychiatry, 50*(9), 1185-1195. (Campbell et al., 2012)

Strengthening Families Program and Life Skills Training #1

Spoth, R., Redmond, C., Trudeau, L., & Shin, C. (2002). Longitudinal substance initiation outcomes for a universal preventive intervention combining family and school programs. *Psychology of Addictive Behaviors, 16*(2), 129–134. https://doi.org/10.1037/0893-164X.16.2.129

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) 372–381. https://doi.org/10.1037/0893-164X.19.4.372

Trudeau, L., Spoth, R., Lillehoj, C., Redmond, C., & Wickrama, K. A. S. (2003). Effects of a preventive intervention on adolescent substance use initiation, expectancies, and refusal intentions. *Prevention Science, 4*, 109-122. https://doi.org/10.1023/a:1022926332514

Spoth, R. L., Randall, G. K., Trudeau, L., Shin, C., & Redmond, C. (2008). Substance use outcomes 5½ years past baseline for partnership-based, family-school preventive interventions. *Drug and Alcohol Dependence, 96*(1-2) 57-68. https://doi.org/10.1016/j.drugalcdep.2008.01.023

Life Skills Training

Luna-Adame, M., Carrasco-Giménez, T., & del Mar Rueda-García, M. (2013). Evaluation of the effectiveness of a smoking prevention program based on the 'Life Skills Training' approach. *Health Education Research, 28*(4), 673-682. https://doi.org/10.1093/her/cyt061

Optimism and Lifeskills Program

Quayle, D., Dziurawiec, S., Roberts, C., Kane, R., & Ebsworthy, G. (2001). The effect of an optimism and lifeskills program on depressive symptoms in preadolescence. *Behavior Change, 18*(4), 194-203. https://doi.org/10.1375/bech.18.4.194

Aussie Optimism Program

Roberts, C. M., Kane, R., Bishop, B., Cross, D., Fenton, J., & Hart, B. (2010). The prevention of anxiety and depression in children from disadvantaged schools. *Behavior Research and Therapy, 48*(1), 68-73. https://doi.org/10.1016/j.brat.2009.09.002

Positive Thinking Program

Rooney, R., Roberts, C., Kane, R., Pike, L., Winsor, A., White, J., & Brown, A. (2006). The prevention of depression in 8- to 9-year-old children: A pilot study. Australian *Journal of Psychologists and Counsellors in Schools, 16*(1), 76-90. https://doi.org/10.1375/ajgc.16.1.76

Emotional Intelligence Training Program

Ruiz-Aranda, D., Castillo, R., Salguero, J. M., Cabello, R., Fernández-Berrocal, P., & Balluerka, N. (2012). Short- and mid-term effects of emotional intelligence training on adolescent mental health. *Journal of Adolescence Health, 51*(5), 462–467. https://dx.doi.org/10.1016/j.jadohealth.2012.02.003.

Tools for Getting Along

Daunic, A. P., Smith, S. W., Garvan, C. W., Barber, B. R., Becker, M. K., Peters, C. D., Taylor, G. G., Van Loan, C. L., Li, W., & Naranjo, A. H. (2012). Reducing developmental risk for emotional/behavioral problems: A randomized controlled trial examining the Tools for Getting Along curriculum. *Journal of School Psychology, 50*(2), 149-166. https://doi.org/10.1016/j.jsp.2011.09.003

Smith, S. W., Daunic, A. P., Barber, B. R., Aydin, B., Van Loan, C. L., & Taylor, G. G. (2014). Preventing risk for significant behavior problems through a cognitive-behavioral intervention: effects of the Tools for Getting Along curriculum at one-year follow-up. *The Journal of Primary Prevention, 35*(5), 371-387. https://doi.org/10.1007/s10935-014-0357-0

Think Smart

Johnson, K. W., Shamblen, S., Ogilvie, K. A., Collins, D., & Saylor, B. (2009). Preventing youths' use of inhalants and other harmful legal products in frontier Alaskan communities: A randomized trial. *Prevention Science, 10*, 298-312. https://doi.org/10.1007/s11121-009-0132-2

RAP-Kiwi; Resourceful Adolescent Program (RAP)

Merry, S., McDowell, H., Wild, C. J., Julliet, B., & Cunliffe, R. (2004). A randomized placebo-controlled trial of a school-based depression prevention program. *Journal of American Academy of Child & Adolescent Psychiatry, 43*(5), 538-547. https://doi.org/10.1097/00004583-200405000-00007

Ease of Handling Social Aspects in Everyday of Life Training

Pössel, P., Horn, A.B., Goren, G., Hautzinger, M. (2004). School-based prevention of depressive symptoms in adolescents: a 6-month follow-up. *Journal of the American Academy of Child and Adolescent Psychiatry, 43*(8), 1003-1010. https://doi.org/10.1097/01.chi.0000126975.56955.98

Cognitive and Social Skills Program

Sarason, I. G., & Sarason, B. R. (1981). Teaching cognitive and social skills to high school students. *Journal of Consulting and Clinical Psychology, 49*(6), 908–919. https://doi.org/10.1037/0022-006X.49.6.908

Roots of Empathy

Santos, R. G., Chartier, M. J., Whalen, J. C., Chateau, D., & Boyd, L. (2011). Effectiveness of school-based violence prevention for children and youth: Cluster randomized field trial of the Roots of Empathy program with replication and three-year follow-up. *Healthcare Quarterly, 14*(2), 80-91. https://doi.org/10.12927/hcq.2011.22367

Problem Solving for Life

Spence, S. H., Sheffield, J., & Donovan, C. L. (2003). Preventing adolescent depression: An evaluation of the Problem Solving for Life program. *Journal of Consulting and Clinical Psychology, 71*(1), 3–13. https://doi.org/10.1037/0022-006X.71.1.3

Spence, S. H., Sheffield, J. K., & Donovan, C. L. (2005). Long-term outcome of a school-based, universal approach to prevention of depression in adolescents. *Journal of Consulting and Clinical Psychology, 73*(1), 160-167. https://doi.org/10.1037/0022-006X.73.1.160

Beyondblue Schools Research Initiative

Sawyer, M. G., Pfeiffer, S., Spence, S. H., Bond, L., Graetz, B., Kay, D., Patton, G., Sheffield, J. (2010). School-based prevention of depression: A randomized controlled study of the beyondblue schools research initiative. *Journal of Child Psychology and Psychiatry, 51*(2), 199-209. https://doi.org/10.1111/j.1469-7610.2009.02136.x

Sawyer, M. G., Harchak, T. F., Spence, S. H., Bond, L., Graetz, B., Kay, D., Patton, G., & Sheffield, J. (2010). School-based prevention of depression: A 2-year follow-up of a randomized controlled trial of the beyondblue schools research initiative. *Journal Of Adolescent Health, 47*(3), 297-304. doi:10.1016/j.jadohealth.2010.02.007

Reading Recovery

Center, Y., Wheldall, K., Freeman, L., Outhred, L., & McNaught, M. (1995). An evaluation of reading recovery. *Reading Research Quarterly, 30*(2), 240–263. https://doi.org/10.2307/748034

Strengthening Families Program and Life Skills Training #2

Spoth, R., Redmond, C., Shin, C., Greenberg, M., Clair, S., Feinberg, M., 2007c. Substance use outcomes at 18 months past baseline: the PROSPER community–university partnership trial. *American Journal of Preventive Medicine, 32*(5), 395–402. https://doi.org/10.1016/j.amepre.2007.01.014

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*(5), 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., Redmond, C., Shin, C., Greenberg, M., Feinberg, M., & Schainker, L. (2013). PROSPER community–university partnership delivery system effects on substance misuse through 6 ½ years past baseline from a cluster randomized controlled intervention trial. *Preventive Medicine, 56*(3-4), 190-196. https://doi.org/10.1016/j.ypmed.2012.12.013

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

Living Letters #1

van der Kooy-Hofland, V. A. C., van der Kooy, J., Bus, A. G., van IJzendoorn, M. H., & Bonsel, G. J. (2012). Differential susceptibility to early literacy intervention in children with mild perinatal adversities: Short- and long-term effects of a randomized control trial. *Journal of Educational Psychology, 104*(2), 337–349. https://doi.org/10.1037/a0026984

Living Letters #2

Clarke, P. J., Snowling, M. J., Truelove, E., & Hulme, C. (2010). Ameliorating children's reading comprehension difficulties: A randomized controlled trial. *Psychological Science, 21*(8), 1106–1116. https://doi.org/10.1177/0956797610375449

Computer-Assisted Learning Program

Ecalle, J., Magnan, A., & Calmus, C. (2009). Lasting effects on literacy skills with a computer-assisted learning using syllabic units in low progress readers. *Computers and Education, 52*(3), 554–561. https://doi.org/10.1016/j.compedu.2008.10.010

Phonemic Awareness and Letter Sound Training

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

Omega-Interactive Sentences, Computerized Phonological Training

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

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

Phonological Awareness Training

Kozminsky, L., & Kozminsky, E. (1995). The effects of early phonological awareness training on reading success. *Learning and Instruction, 5*(3), 187-201. https://doi.org/10.1016/0959-4752(95)00004-M

Morphological versus Phonological Awareness Training

Lyster, S. A. H. (2002). The effects of morphological versus phonological awareness training in kindergarten on reading development. *Reading and Writing, 15*(3-4), 261-294. https://doi.org/10.1023/A:1015272516220

Lyster, S. A. H., Lervåg, A. O., & Hulme, C. (2016). Preschool morphological training produces long-term improvements in reading comprehension. *Reading and Writing, 29*(6), 1269-1288. https://doi.org/10.1007/s11145-016-9636-x

School-to-Jobs/Pathways for Youth

Oyserman, D., Bybee, D., & Terry, K. (2006). Possible selves and academic outcomes: How and when possible selves impel action. *Journal of Personality and Social Psychology, 91*(1), 188-204. https://doi.org/10.1037/0022-3514.91.1.188

Computer-Assisted Reading Intervention for Children at Risk of Dyslexia

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

Reading Summer Day-Camp

Schacter, J., & Jo, B. (2005). Learning when school is not in session: a reading summer day-camp intervention to improve the achievement of exiting first-grade students who are economically disadvantaged. *Journal of Research in Reading, 28*(2), 158-169. https://doi.org/10.1111/j.1467-9817.2005.00260.x

Reading Remediation for Children with Reading Disorders

Gittelman, R., & Feingold, I. (1983). Children with reading disorders: Efficacy of reading remediation. *Journal of Child Psychology & Psychiatry, 24*(2), 167–191. https://doi.org/10.1111/j.1469-7610.1983.tb00568.x

Read Well Kindergarten

Gunn, B., Smolkowski, K., & Vadasy, P. (2010). Evaluating the effectiveness of Read Well Kindergarten. *Journal of Research on Educational Effectiveness, 4*(1), 53–86. https://doi.org/10.1080/19345747.2010.488716

Multi-Component Reading Remediation

Morris, R. D., Lovett, M. W., Wolf, M., Sevcik, R. A., Steinbach, K. A., Frijters, J. C., & Shapiro, M. B. (2012). Multiple-component remediation for developmental reading disabilities: IQ, socioeconomic status, and race as factors in remedial outcome. *Journal of Learning Disabilities, 45*(2), 99-127. https://doi.org/10.1177/0022219409355472

Early Literacy Concepts

Phillips, L. M., Norris, S. P., & Mason, J. M. (1996). Longitudinal effects of early literacy concepts on reading achievement: A kindergarten intervention and five-year follow-up. *Journal of Literacy Research, 28*(1), 173-195. https://doi.org/10.1080/10862969609547915

Supplemental Phonics-based Instruction

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

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

Phonological/Early Reading Skills

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

Code-Oriented Reading Instruction

Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2006). Code-oriented instruction for kindergarten students at risk for reading difficulties: A randomized field trial with paraeducator implementers. *Journal of Educational Psychology, 98*(3), 508–528. https://doi.org/10.1037/0022-0663.98.3.508

Computer-Assisted Blending Skill Training

Reitsma, P., & Wesseling, R. (1998). Effects of computer-assisted training of blending skills in kindergarteners. *Scientific Studies of Reading, 2*(4), 301–320. https://doi.org/10.1207/s1532799xssr0204_1

Cogmed Working Memory Training

Roberts, G., Quach, J., Spencer-Smith, M., Anderson, P. J., Gathercole, S., Gold, L., Sia, K., Mensah, F., Rickards, F., Ainley, J., & Wake, M. (2016). Academic outcomes 2 years after working memory training for children with low working memory: A randomized clinical trial. *JAMA Pediatrics, 170*(5) 1-10. https://doi.org/10.1001/jamapediatrics.2015.4568

Explicit Phonological Awareness Instruction

Ryder, J. F., Tunmer, W. E., & Greaney, K. T. (2008). Explicit instruction in phonemic awareness and phonemically based decoding skills as an intervention strategy for

struggling readers in whole language classrooms. *Reading and Writing, 21*(4), 349–369. https://doi.org/10.1007/s11145-007-9080-z

Project Care

Wasik, B.H., Ramey, C.T., Bryant, D.M., Sparling, J.J. (1990). A longitudinal study of two early intervention strategies: Project CARE. *Child Development, 61*(6), 1682-1696. https://doi.org/10.1111/j.1467-8624.1990.tb03559.x

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

LARS & LISA (Lust An Realistischer Sicht & Leichtigkeit Im Sozialen Alltag)

Pössel, P., Adelson, J. L., & Hautzinger, M. (2011). A randomized trial to evaluate the course of effects of a program to prevent adolescent depressive symptoms over 12 months. *Behaviour Research and Therapy, 49*(12), 838-851. https://doi.org/10.1016/j.brat.2011.09.010

Pössel, P., Seemann, S., & Hautzinger, M. (2008). Impact of comorbidity in prevention of adolescent depressive symptoms. *Journal of Counseling Psychology, 55(1), 106–117.* https://doi.org/10.1037/0022-0167.55.1.106

Consultation-based Intervention

DuPaul, G. J., Jitendra, A. K., Volpe, R. J., Tresco, K. E., Lutz, J. G., Vile Junod, R. E., Cleary, K. S., Flammer, L. M., Mannella, M. C. (2006) Consultation-based academic interventions for children with ADHD: effects on reading and mathematics achievement. *Journal of Abnormal Child Psychology*, *34*(5), 635-48. https://doi.org/10.1007/s10802-006-9046-7

Volpe, R. J., DuPaul, G. J., Jitendra, A. K., & Tresco, K. E. (2009). Consultation-based academic interventions for children with attention deficit hyperactivity disorder: effects on reading and mathematics outcomes at 1-year follow-up. *School Psychology Review, 38*(1), 5-13.

Teacher Responsivity Education

Cabell, S. Q., Justice, L. M., Piasta, S. B., Curenton, S. M., Wiggins, A., Turnbull, K. P., & Petscher, Y. (2011). The impact of teacher responsivity education on preschoolers' language and literacy skills. *American Journal of Speech-Language Pathology, 20*(4), 315–330. https://doi.org/10.1044/1058-0360(2011/10-0104)

Johanson, M., Justice, L. M., & Logan, J. (2015). Kindergarten impacts of a preschool language-focused intervention. *Applied Developmental Science, 20*(2), 94-107. https://doi-org.tc.idm.oclc.org/10.1080/10888691.2015.1074050

Head Start Research-Based, Developmentally-Informed Program

Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A., Greenberg, M. T., Blair, C., Nelson, K. E.,. Gill, S. (2008). Promoting academic and social-emotional school readiness: The head start REDI program. *Child Development, 79*(6), 1802-1817. https://doi.org/10.1111/j.1467-8624.2008.01227.x

Bierman, K. L., Nix, R. L., Heinrichs, B. S., Domitrovich, C. E., Gest, S. D., Welsh, J. A., & Gill, S. (2014). Effects of Head Start REDI on children's outcomes 1 year later in different kindergarten contexts. *Child Development, 85*(1), 140–159. https://doi.org/10.1111/cdev.12117

Nix, R. L., Bierman, K. L., Heinrichs, B. S., Gest, S. D., Welsh, J. A., & Domitrovich, C. E. (2016). The randomized controlled trial of Head Start REDI: Sustained effects on

developmental trajectories of social–emotional functioning. *Journal of Consulting and Clinical Psychology, 84*(4), 310–322. https://doi.org/10.1037/a0039937

Bierman, K. L., Heinrichs, B. S., Welsh, J. A., & Nix, R. L. (2021). Reducing adolescent psychopathology in socioeconomically disadvantaged children with a preschool intervention: A randomized controlled trial. *American Journal of Psychiatry, 178*(4), 305-312. https://doi.org/10.1176/appi.ajp.2020.20030343

Safe Dates

Foshee, V. A., Bauman, K. E., Arriaga, X. B., Helms, R. W., Koch, G. G., & Linder, G. F. (1998). An evaluation of safe dates, an adolescent dating and violence prevention program. *American Journal of Public Health, 88*(1), 45-50. https://doi.org/10.2105/AJPH.88.1.45

Foshee, V. A., Bauman, K. E., Greene, W. F., Koch, G. G., Linder, G. F., & MacDougall, J. E. (2000). The Safe Dates Program: 1-year follow-up results. *American Journal of Public Health, 90*(10), 1619-1622. https://doi.org/10.2105/ajph.90.10.1619

Foshee, V.A., Bauman, K.E., Ennett, S.T., Suchindran, C., Benefield, T., & Linder, G.F. (2005). Assessing the effects of the dating violence prevention program "Safe Dates" using random coefficient regression modeling. *Prevention Science, 6*, 245-258. https://doi.org/10.1007/s11121-005-0007-0

SEARCH Screening Test and TEACH Tutoring

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

English Reading Intervention for English Language Learners

Vaughn, S., Cirino, P. T., Linan-Thompson, S., Mathes, P. G., Carlson, C. D., Hagan, E. C., Pollard-Durodola, S. D., Fletcher, J. M., & Francis, D. J. (2006). Effectiveness of a Spanish intervention and an English intervention for English-language learners at risk for reading problems. *American Educational Research Journal, 43*(3), 449–487.

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

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

Spanish Reading Intervention for English Language Learners

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

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

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

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

Swedish Phonics-based Intervention

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

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

Computer-Assisted Remedial Reading Intervention

Saine, N. L., Lerkkanen, M., Ahonen, T., Tolvanen, A., & Lyytinen, H. (2010). Predicting word-level reading fluency outcomes in three contrastive groups: Remedial and computer-assisted remedial reading intervention, and mainstream instruction. *Learning and Individual Differences, 20*(5), 402–414. https://doi.org/10.1016/j.lindif.2010.06.004

Home-Based Dyslexia Prevention

van Otterloo, S. G., van der Leij, A., & Henrichs, L. F. (2009). Early home-based intervention in the Netherlands for children at familial risk of dyslexia. *Dyslexia: An International Journal of Research and Practice, 15*(3), 187–217. https://doi.org/10.1002/dys.376

Supplemental Reading Instruction

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

Word Analysis Skills Training

Lie, A. (1991). Effects of a training program for stimulating skills in word analysis in first-grade children. *Reading Research Quarterly, 26*(3), 234-250. https://doi.org/10.2307/747762

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

Childhood Depression Prevention

Cecchini, T. B. (1997). *An interpersonal and cognitive-behavioral approach to childhood depression: A school-based primary prevention study* (Order No. 9820698). Available from ProQuest Dissertations & Theses Global. (304420573).

http://ezproxy.cul.columbia.edu/login?url=https://www.proquest.com/dissertations-theses/interpersonal-cognitive-behavioral-approach/docview/304420573/se-2

Johnson, N. C. (2000). *A follow-up study of a primary prevention program targeting childhood depression* (Order No. 1402700). Available from ProQuest Dissertations & Theses Global. (304634088). http://ezproxy.cul.columbia.edu/login?url=https://www.proquest.com/dissertations-theses/follow-up-study-primary-prevention-program/docview/304634088/se-2

FRIENDS Program #2

Lowry-Webster, H. M., Barrett, P. M., & Dadds, M. R. (2001). A universal prevention trial of anxiety and depressive symptomatology in childhood; Preliminary data from an Australian study. *Behaviour Change, 18*(1), 36-50. https://doi.org/10.1375/bech.18.1.36

Lowry-Webster, H. M., Barrett, P. M., & Lock, S. (2003). A universal prevention trial of anxiety symptomatology during childhood; Results at 1-year follow-up. *Behaviour Change, 20*(1), 25-43. https://doi.org/10.1375/bech.20.1.25.24843

FRIENDS Program #4

Essau, C. A., Conradt, J., Sasagawa, S., & Ollendick, T. H. (2012). Prevention of anxiety symptoms in children: Results from a universal school-based trial. *Behavioral Therapy, 43*(2), 450-464. https://doi.org/10.1016/j.beth.2011.08.003

Life Skills Training in Minority Youth

Botvin, G. J., Griffin, K. W., Diaz, T., & Ifill-Williams, M. (2001). Preventing binge drinking during early adolescence: One- and two-year follow-up of a school-based preventive intervention. *Psychology of Addictive Behaviors, 15*(4), 360-365. https://doi.org/10.1037/0893-164X.15.4.360

Classroom-Centered and School-Family Partnership Interventions

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

Ialongo, N., Poduska, J., Werthamer, L., & Kellam, S. (2001). The distal impact of two first-grade preventive interventions on conduct problems and disorder in early adolescence. *Journal of Emotional & Behavioral Disorders, 9*(3), 146-161.

Storr, C. L., Ialongo, N. S., Kellam, S. G., & Anthony, J. C., (2002). A randomized controlled trial of two primary school intervention strategies to prevent early onset tobacco smoking. *Drug and Alcohol Dependence, 66*(1), 51–60. https://doi.org/10.1016/S0376-8716(01)00184-3

Furr-Holden, C. D., Ialongo, N. S., Anthony, J. C., Petras, H., Kellam, S. G., (2004). Developmentally inspired drug prevention: middle school outcomes in a school-based randomized prevention trial. *Drug and Alcohol Dependence, 73*(2), 149–158. https://doi.org/10.1016/j.drugalcdep.2003.10.002

Wang, Y., Browne, D. C., Petras, H., Stuart, E. A., Wagner, F. A., Lambert, S. F., Kellam, S. G., & Ialongo, N. S. (2009). Depressed mood and the effect of two universal first grade preventive interventions on survival to the first tobacco cigarette smoked among urban youth. *Drug and Alcohol Dependence, 100*(3), 194-203. https://doi.org/10.1016/j.drugalcdep.2008.08.020

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), 926-927. https://doi.org/10.1037/a0016586

Parent-Adolescent Interaction

Dunham, R., Kidwell, J., & Portes, P. (1988). Effects of parent-adolescent interaction on the continuity of cognitive development from early childhood to early adolescence. *Journal of Early Adolescence, 8*(3), 297-310. https://doi-org.tc.idm.oclc.org/10.1177/0272431688083006

Infant Health and Development Program

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

McCarton, Brooks-Gunn, Wallace, et al. (1997). Results at age 8 years of early intervention for low-birth-weight premature infants: The infant health and development program. *The Journal of the American Medical Association, 277*(2), 126-132. https://doi.org/10.1001/jama.1997.03540260040033

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

Classroom and At-Home Preschool Interventions

U.S. Department of Health, Education, and Welfare Office of Education. (1969). *Investigations of classroom and at-home interventions: Research and development program on preschool disadvantaged children* (M. B. Karnes, A. S. Hodgins, J. A. Teska, & S. A. Kirk, Authors; Research Report No. ED 036 663). https://files.eric.ed.gov/fulltext/ED036663.pdf

Penn Resiliency Program #1

Cardemil, E. V., Reivich, K. J., & Seligman, M. E. P. (2002). The prevention of depressive symptoms in low-income minority middle-school students. *Prevention & Treatment, 5*(7). https://doi-org.tc.idm.oclc.org/10.1037/1522-3736.5.1.58a

Penn Resiliency Program #2

Gillham, J. E., Reivich, K. J., Freres, D. R., Chaplin, T. M., Shatté, A. J., Samuels, B., Elkon, A. G. L. Litzinger, S., Lascher, M., Gallop, R., & Seligman, M. E. P. (2007). School-based prevention of depressive symptoms: A randomized controlled study of the effectiveness and specificity of the Penn Resiliency program. *Journal of Consulting and Clinical Psychology, 75*(1), 9-19. https://doi-org.tc.idm.oclc.org/10.1037/0022-006X.75.1.9

Linking the Interests of Families and Teachers

Eddy, J. M., Reid, J. B., Stoolmiller, M., & Fetrow, R. (2003). Outcomes during middle school for an elementary school-based preventive intervention for conduct problems: Follow-up results from a randomized trial. *Behavior Therapy, 34*(4), 535-552. https://doi.org/10.1016/S0005-7894(03)80034-5

Universal School-based Mental Health Intervention

Collins, S., Woolfson, M. L., Durkin, K. (2014). Effects on coping skills and anxiety of a universal school-based mental health intervention delivered in Scottish primary schools.

School Psychology International, 35(1), 85-100. https://doi.org/10.1177/0143034312469157

Unplugged

Faggiano, F., Galanti, M. R., Bohrn, K., Burkhart, G., Vigna-Taglianti, F., Cuomo, L., Fabiani, L., Panella, M., Perez, T., Siliquini, R., van der Kreeft, P., Vassara, M., & Wiborg, G., (2008). The effectiveness of a school-based substance abuse prevention program: EU-Dap cluster randomized controlled trial. *Preventive Medicine, 47*(5), 537-543. https://doi.org/10.1016/j.ypmed.2008.06.018

Faggiano, F., Richardson, C., Bohrn, K., Galanti, M. R., & EU-Dap Study Group (2007). A cluster randomized controlled trial of school-based prevention of tobacco, alcohol and drug use: The EU-Dap design and study population. *Preventive Medicine, 44*(2), 170-173. https://doi.org/10.1016/j.ypmed.2006.09.010

Faggiano, F., Vigna-Taglianti, F., Burkhart, G., Bohrn, K., Cuomo, L., Gregori, D., Panella, M., Scatigna, M., Siliquini, R., Varona, L., van der Kreeft, P., Vassara, M., Wiborg, G., Rosaria Galanti, M., & the EU-Dap Study Group (2010). The effectiveness of a school-based substance abuse prevention program: 18-month follow-up of the EU-Dap cluster randomized controlled trial. *Drug and Alcohol Dependence, 108*(1-2), 56-64. https://doi.org/10.1016/j.drugalcdep.2009.11.018

Responding in Peaceful and Positive Ways (Seventh Grade)

Farrell, A., Meyer, A., Sullivan, T., & Kung, E. (2003). Evaluation of the Responding in Peaceful and Positive Ways Seventh Grade (RIPP-7) universal violence prevention program. *Journal of Child and Family Studies, 12*, 101-120. https://doi.org/10.1023/A:1021314327308

Head Start Classroom-based Approaches and Resources for Emotion and Social Skill Promotion Administration for Children and Families Office of Planning, Research, and Evaluation. (2014). *Impact findings from the Head Start CARES demonstration: National evaluation of three approaches to improving preschoolers' social and emotional competence* (P. Morris, S. K. Mattera, N. Castells, M. Bangser, K. Bierman, & C. Raver, Authors; Research Report No. OPRE 2014-44). U.S. Department of Health and Human Services. https://files.eric.ed.gov/fulltext/ED546649.pdf

Early Head Start

Administration for Children and Families Office of Planning, Research, and Evaluation, Child Outcomes Research and Evaluation & Administration on Children, Youth, and Families, Head Start Bureau. (2002). *Making a difference in the lives of infants and toddlers and their families: The impacts of Early Head Start* (J. M. Love, E. E. Kisker, C. M. Ross, J. Brooks-Gunn, D. Paulsell, K. Boller, . . . C. Brady-Smith, Authors). U.S. Department of Health and Human Services; Mathematica. https://www.acf.hhs.gov/sites/default/files/documents/opre/impacts_vol1.pdf

Administration for Children and Families Office of Planning, Research, and Evaluation. (2010). *Early Head Start children in grade 5: Long-term follow up of the Early Head Start Research and Evaluation Project study sample* (C. A. Vogel, Y. Xue, E. M. Moiduddin, B. L. Carlson, & E. E. Kisker, Authors; Report No. OPRE 2011-8). U.S. Department of Health and Human Services.

Tennessee Pre-K

Lipsey, M. W., Farran, D. C., & Hofer, K. G. (2015). *A randomized control trial of a statewide voluntary prekindergarten program on children's skills and behaviors through third*

grade. [Research Report]. Peabody Research Institute, Vanderbilt University. https://files.eric.ed.gov/fulltext/ED566664.pdf

Lipsey, M. W., Farran, D. C., & Durkin, K. (2018). Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly, 45*(4), 155-176. https://doi.org/10.1016/j.ecresq.2018.03.005

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

Head Start Impact Study

Administration for Children and Families Office of Planning, Research, and Evaluation. (2010). *Head Start Impact Study: Final report* (M. Puma, S. Bell, R. Cook, C. Heid, G. Shapiro, P. Broene, . . . E. Spier, Authors). U.S. Department of Health and Human Services; Westat. https://files.eric.ed.gov/fulltext/ED507845.pdf

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., Mashburn, A., & Downer, J. (2012). *Third grade follow-up to the Head Start Impact Study*. [Final report]. U.S. Administration for Children and Families.

Dialogic Reading #1

Whitehurst, G.J., Arnold, D.S., Epstein, J.N., & Angell, A.L. (1994). A picture book reading intervention in day care and home for children from low-income families. *Developmental Psychology, 30*(5), 679–689. https://doi-org.tc.idm.oclc.org/10.1037/0012-1649.30.5.679

Nurse Home Visitation Program

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

Olds D. L., Henderson, C. R., Cole R., Eckenrode, J., Kitzman, H., Luckey, D., Pettitt, L., Sidora, K., Morris, P., & Powers, J. (1998) Long-term effects of nurse home visitation on children's criminal and antisocial behavior 15-year follow-up of a randomized controlled trial. *JAMA, 280*(14), 1238–1244. https://doi.org/10.1001/jama.280.14.1238

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

FRIENDS Program #1

Barrett, P. M, Lock, S., & Farrell, L. J. (2005). Developmental differences in universal preventive intervention for child anxiety. *Clinical Child Psychology and Psychiatry, 10*(4), 539-555. https://doi.org/10.1177/1359104505056317

Dialogic Reading #2

Whitehurst, G. J., Falco, F. L., Lonigan, C. J., Fischel, J. E., DeBaryshe, B. D., Valdez-Menchaca, M. C., & Caulfield, M. (1988). Accelerating language development through picture book reading. *Developmental Psychology, 24*(4), 552–559. https://doi.org/10.1037/0012-1649.24.4.552

Perry Preschool

Weikart, D. P. (1970). Longitudinal Results of the Ypsilanti Perry Preschool Project. Final Report. Volume II of 2 Volumes.

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

The Early Training Project

Klaus, R. A., & Gray, S. W. (1968) The Early Training Project for disadvantaged children: a report after five years. *Monographs of the Society for Research in Child Development, 33*(4), 1-66. https://doi.org/10.2307/1165717

Gray, S. W., & Klaus, R. A. (1970). The Early Training Project: A seventh-year report. *Child Development, 41*(4), 909-924. https://doi.org/10.2307/1127321

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

Phonics-Based Instruction for First Graders

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

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

Parent Training for Teenage Moms

Field, T., Widmayer, S., Greenberg, R., Stoller, S. (1982). Effects of parent training on teenage mothers and their infants. *Pediatrics, 69*(6), 703-707. https://doi.org/10.1542/peds.69.6.703

Abecedarian

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. (2016). *Early childhood education.* NBER 2.pdf

Campbell, F. A., & Ramey, C. T. (1995). Cognitive and school outcomes for high-risk African-American students at middle adolescence: Positive effects of early intervention. *American Educational Research Journal, 32*(4), 743-772. https://doi.org/10.3102/00028312032004743

Campbell, F. A., Pungello, E. P., Burchinal, M., Kainz, K., Pan, Y., Wasik, B. H., Barbarin, O., A., Sparling, J., J., & Ramey, C. T. (2012). Adult outcomes as a function of an early childhood educational program: an Abecedarian Project follow-up. *Developmental Psychology, 48*(4), 1033. https://doi.org/10.1037/a0026644

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education/ Young adult outcomes from the Abecedarian Project. *Applied developmental science, 6*(1), 42-57. https://doi.org/10.1207/S1532480XADS0601_05

Campbell, F.A., & Ramey, C.T. (1994). Effects of early intervention on intellectual and academic achievement: A follow-up study of children from low-income families. *Child Development, 65*(2), 684–698. https://doi.org/10.1111/j.1467-8624.1994.tb00777.x

Technology-Enhanced, Research-Based, Instruction, Assessment, and Professional Development Scale-up Model

Sarama, J., Lange, A. A., Clements, D. H., & Wolfe, C. B. (2012). The impacts of an early mathematics curriculum on oral language and literacy. *Early Childhood Research Quarterly, 27*(3), 489-502. https://doi.org/10.1016/j.ecresq.2011.12.002

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

Reading with Rhyme, Reading with Phoneme

Hatcher, P. J., Hulme, C., & Snowling, M. J. (2004). Explicit phoneme training combined with phonic reading instruction helps young children at risk of reading failure. *Journal of Child Psychiatry, 45*(2), 338–358. https://doi.org/10.1111/j.1469-7610.2004.00225.x

FRIENDS Program #3

Lock, S. & Barrett, P. M. (2003). A longitudinal study of developmental differences in universal preventive intervention for child anxiety. *Behaviour Change, 20*(4), 183-199. https://doi.org/10.1375/bech.20.4.183.29383

Barrett, P. M., Farrell, L. J., Ollendick, T. H., & Dadds, M. (2006). Long-term outcomes of an Australian universal prevention trial of anxiety and depression symptoms in children and youth: An evaluation of the Friends program. *Journal of Clinical Child & Adolescent Psychology, 35*(3) 403-411. https://doi.org/10.1207/s15374424jccp3503_5

Cognitive-Behavioral Approach to Drug Abuse Prevention

Botvin, G. J., Baker, E., Dusenbury, L., Tortu, S., & Botvin, E. M. (1990). Preventing adolescent drug abuse through a multimodal cognitive-behavioral approach: Results of a 3-year study. *Journal of Consulting and Clinical Psychology, 58*(4), 437-446. https://doi.org/10.1037/0022-006X.58.4.437

Botvin, G. J., Baker, E., Dusenbury, L., Botvin, E. M., & Diaz, T. (1995). Long-term follow-up results of a randomized drug abuse prevention trial in a white middle-class population. *The Journal of the American Medical Association, 273*(14), 1106-1112. https://doi.org/10.1001/jama.1995.03520380042033

Botvin, G. J., Griffin, K. W., Diaz, T., Scheier, L. M., Williams, C., & Epstein, J. A. (2000). Preventing illicit drug use in adolescents: Long-term follow-up data from a randomized control trial of a school population. *Addictive Behaviors, 25*(5), 769-774. https://doi.org/10.1016/S0306-4603(99)00050-7

Griffin, K. W., Botvin, G. J., & Nichols, T. R. (2006). Effects of a school-based drug abuse prevention program for adolescents on HIV risk behavior in young adulthood. *Prevention Science, 7*, 103-112. https://doi.org/10.1007/s11121-006-0025-6

Peer-Led Life Skills Training

Botvin, G. J., Baker, E., Renick, N. L., Filazzola, A. D., & Botvin, E. M. (1984).. *Addictive Behaviors, 9*(2), 137-147. https://doi.org/10.1016/0306-4603(84)90051-0

Botvin, G. J., Baker, E., Filazzola, A. D., & Botvin, E. M. (1990). A cognitive-behavioral approach to substance abuse prevention: One-year follow-up. *Addictive Behaviors, 15*(1), 47-63. https://doi.org/10.1016/0306-4603(90)90006-J

Culturally-Focused Substance Use Prevention

Botvin, G. J., Schinke, S. P., Epstein, J. A., & Diaz, T. (1994). Effectiveness of culturally focused and generic skills training approaches to alcohol and drug abuse prevention among minority youths. *Psychology of Addictive Behaviors, 8*(2), 116-127. https://doi.org/10.1037/0893-164X.8.2.116

Botvin, G. J., Schinke, S. P., Epstein, J. A., Diaz, T., & Botvin, E. M. (1995). Effectiveness of culturally focused and generic skills training approaches to alcohol and drug abuse prevention among minority adolescents: Two-year follow-up results. *Psychology of Addictive Behaviors, 9*(3), 183–194. https://doi.org/10.1037/0893-164X.9.3.183

Second Step: A Violence Prevention Curriculum

Grossman, D. C., Neckerman, H. J., Koepsell, T. D., Liu, P. Y., & Asher, K. N. (1997). Effectiveness of a violence prevention curriculum among children in elementary school. *Journal of the American Medical Association*, *277*(20), 1505-1611. https://doi.org/10.1001/jama.1997.03540440039030

Conflict-Resolution Training Program

Güneri, O. Y., & Çoban, R. (2004). The effect of conflict resolution training on Turkish elementary school students: A quasi-experimental investigation. International *Journal for the Advancement of Counselling, 26*, 109-124. https://doi.org/10.1023/B:ADCO.0000027425.24225.87

All Stars

Harrington, N. G., Giles, S. M., Hoyle, R. H., Feeney, G. J., & Yungbluth, S. C. (2001). Evaluation of the All Stars character education and problem behavior prevention program: Effects on mediator and outcome variables for middle school students. *Health Education and Behavior, 28*(5), 533-546. https://doi.org/10.1177/109019810102800502

Eigenständig Werden (Fifth and Sixth Grade)

Hansen, J., Hanewinkel, R., Maruska, K., & Isensee, B. (2011) The 'Eigenständig werden' prevention trial: a cluster randomised controlled study on a school-based life skills programme to prevent substance use onset. *BMJ Open 1*(2), e000352. https://doi.org/10.1136/bmjopen-2011-000352

Isensee, B., Hansen, J., Maruska, K., & Hanewinkel, R. (2014). Effects of a school-based prevention programme on smoking in early adolescence: a 6-month follow-up of the 'Eigenständig werden'cluster randomised trial. *BMJ open, 4*(1), e004422. https://doi.org/10.1136/bmjopen-2013-004422