EdWorkingPaper No. 23-885

Are Preschool Programs Becoming Less Effective?

Anamarie A. Whitaker
University of Delaware

Margaret Burchinal
University of Virginia

Jade M. Jenkins
University of California, Irvine

Drew H. Bailey
University of California, Irvine

Tyler W. Watts
Teachers College, Columbia University

Greg J. Duncan
University of California, Irvine

Emma R. Hart
Teachers College, Columbia University

Ellen Peisner-Feinberg
University of North Carolina

High-quality preschool programs are heralded as an effective policy tool to promote the development and life-long wellbeing of children from low-income families. Yet evaluations of recent preschool programs produce puzzling findings, including negative impacts, and divergent, weaker results than were shown in demonstration programs implemented in the 1960s and 70s. We provide potential explanations for why modern evaluations of preschool programs have produced less positive and more mixed results, focusing on changes in counterfactual conditions and preschool instructional practices. We also address popular explanations such as subsequent low-quality schooling experiences that, we argue, do not appear to account for weakening program effectiveness. The field must take seriously the smaller positive, null, and negative impacts from modern programs and strive to understand why effects vary and how to boost program effectiveness through rigorous, longitudinal research.

VERSION: October 2024

Suggested citation: Whitaker, Anamarie A., Margaret Burchinal, Jade M. Jenkins, Drew H. Bailey, Tyler W. Watts, Greg J. Duncan, Emma R. Hart, and Ellen S. Peisner-Feinberg. (2023). Why are Preschool Programs Becoming Less Effective?. (EdWorkingPaper: 23-885). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/smqa-n695

Are Preschool Programs Becoming Less Effective?

Anamarie A. Whitaker
University of Delaware

Margaret Burchinal
University of Virginia

Jade M. Jenkins
Drew H. Bailey
University of California, Irvine

Tyler W. Watts
Teachers College, Columbia University

Greg J. Duncan
University of California, Irvine

Emma R. Hart
Teachers College, Columbia University

Ellen Peisner-Feinberg
University of North Carolina

Author Note

Anamarie A. Whitaker https://orcid.org/0000-0002-6865-5850

Margaret Burchinal https://orcid.org/0000-0002-3606-7843

Jade M. Jenkins https://orcid.org/0000-0002-2000-3087

Tyler W. Watts https://orcid.org/0000-0002-2741-0873

Greg J. Duncan https://orcid.org/0000-0002-9869-6311

Emma R. Hart https://orcid.org/0000-0003-3808-0838

Ellen Peisner-Feinberg https://orcid.org/0000-0001-9716-9977

We have no conflicts of interests to disclose.

Correspondence concerning this article should be addressed to Anamarie A. Whitaker, 212 Alison Hall, Newark, DE 19716. Email: awhit@udel.edu

Abstract

High-quality preschool programs are heralded as an effective policy tool to promote the development and life-long wellbeing of children from low-income families. Yet evaluations of recent preschool programs produce puzzling findings, including negative impacts, and divergent, weaker results than were shown in demonstration programs implemented in the 1960s and 70s. We provide potential explanations for why modern evaluations of preschool programs have produced less positive and more mixed results, focusing on changes in counterfactual conditions and preschool instructional practices. We also address popular explanations such as subsequent low-quality schooling experiences that, we argue, do not appear to account for weakening program effectiveness. The field must take seriously the smaller positive, null, and negative impacts from modern programs and strive to understand why effects vary and how to boost program effectiveness through rigorous, longitudinal research.

Are preschool programs becoming less effective?

Evidence from small-scale randomized control trials (RCTs) conducted decades ago demonstrated that early childhood educational programs can support young children from low-income families in developing the skills needed to succeed in school and adult life (Elango et al., 2016). These studies, plus quasi-experimental evaluations showing long-run benefits to early cohorts of the Head Start program (e.g., Bailey et al., 2021), have convinced many people that preschool could be one of the most effective policy levers for equalizing opportunity in the United States. They are often cited to justify large investments by federal, state, and local governments in preschool programs for families with low incomes (White House, 2023).

At the same time, randomized evaluations of more recent preschool programs have not replicated the longer-term results of the early evaluations (Burchinal et al., 2024). Although some studies have found impressive positive impacts into young adulthood, others have shown no impacts, and one found significantly lower school achievement and worse behavior among children attending state-funded preschool programs when compared with children who did not. As public spending on preschool increases, it is important to understand why some evaluations produce much more promising results than others.

Like K-12 education, public investments in early education are a central component of a nation's social policy. We do not contest whether governments should be investing in preschool. Rather, our focus is on how to make those investments as effective as possible. We do this by reviewing the evidence on the theoretical underpinnings of commonly promoted explanations for why the longer-term effects of preschool (i.e., effects found one or more years after program completion) are smaller and more uncertain today than they were for earlier preschool programs. Specifically, we identify changes in the counterfactual conditions as one of the most plausible explanations, although we also review evidence on theories regarding instructional focus and subsequent environmental experiences.

More broadly, we present our thesis that modern day preschool programs less consistently impart the long-term advantages than those of early demonstration programs, despite their success in imparting short-term impacts. Many factors likely explain why this is true; there does not exist one piece of unimpeachable evidence, nor a statistical test, to parsimoniously resolve why preschool programs may be becoming less effective in the long-term. However, we argue that this tapestry of evidence lacking a developmental theory to reconcile the findings should motivate all stakeholders to develop the science around long-run impacts. We conclude with recommendations for future research to help build, refine, and test theories of program effectiveness.

Effects of Preschool Participation

RCT evaluations of the researcher-designed Perry Preschool and Abecedarian projects, which operated in the 1960s and 70s, respectively, and served about 60 children, demonstrate that preschool can provide lifetime benefits for disadvantaged children, including significant initial and sustained boosts in achievement scores and, in Abecedarian, IQ (Elango et al., 2016; García et al., 2021). Long-run evaluations of these programs show increased adult education, employment, earnings, and health, and lower levels of crime and incarceration in treatment group compared with control group children, with estimated program-generated benefits totaling six to ten times the costs (García et al., 2021). Evidence from Perry Preschool even suggests spillover benefits to children of the original preschool participants in terms of higher quality family environments and better educational and employment outcomes (García et al., 2023). Similarly, quasi-experimental evaluations of early cohorts of Head Start show long-term positive benefits on educational and employment outcomes (Bailey et al., 2021; Deming, 2009), and provide some evidence of spillover benefits for younger siblings (Garces et al., 2002). In short, results from rigorous evaluations of preschool programs operating approximately 30-50 years ago provide convincing evidence that preschool programs can produce lasting benefits for children from low-income families by improving educational attainment, employment, and health in adulthood.

Recent studies of modern-day, large-scale preschool programs produce a more complicated and inconsistent pattern of results. Reviews of this literature largely suggest positive end of program impacts and mixed longer-term impacts on children's achievement and social emotional development (Burchinal et al., 2024; Cascio, 2021; Phillips et al., 2017). Specifically, evidence from preschool program evaluations tend to suggest moderate to large initial positive achievement associations that decline and sometimes, but not always, completely fadeout during the first years of elementary school followed by inconsistent long-run results (Abenavoli, 2019; Ansari et al., 2020; Logan et al., 2024).

For instance, a regression discontinuity design study of Tulsa public preschool options (pre-kindergarten [pre-k] and Head Start) found positive and significant end of treatment effects for attendees, with pre-k producing larger early literacy gains but not math gains compared with Head Start (Gormley et al., 2008). Longer-term evaluations, which relied on propensity score matching techniques, show positive high school educational outcomes in terms of reduced absences for both preschool options, but significant increases in course-taking and reductions in grade retention for pre-k participants only (Amadon et al., 2022). Further, no significant achievement effects emerged in high school for either preschool option. On the other hand, a propensity score study of the effectiveness of New Jersey pre-k found sustained effects of participation on academic achievement outcomes into high school, particularly for participants with two years of preschool (Barnett & Jung, 2021).

High-quality lottery and RCT designs have the highest degree of internal validity because they can directly control child selection into a program. Perry Preschool and the Abecedarian program evaluations used these designs (Elango et al., 2016). Recent program evaluations using these methods have produced markedly inconsistent results. The nationwide RCT evaluation of the Head Start program yielded some positive impacts on language and literacy at the end of the Head Start year, but these differences disappeared rapidly after children entered elementary school (Puma et al., 2012). By third grade, virtually no statistically significant results were detected. In a recent Boston pre-k evaluation, researchers found positive high school and college going outcomes for lottery winners compared with losers (Gray-Lobe et al., 2023). Interestingly, neither Gray-Lobe et al. nor a lottery-based evaluation of later pre-k attendees in Boston found that pre-k participation benefitted academic skills during the intermediary school years (Gray-Lobe et al., 2023, Weiland et al., 2020). Most concerning, a randomized evaluation of the Tennessee pre-k program found positive end of preschool academic outcomes followed by adverse impacts on academic achievement and school disciplinary outcomes in elementary and middle school (Durkin et al., 2022). Disparate findings such as these are hard to reconcile. They also raise important questions about what medium- to long-term impacts should be expected from today's preschool programs and how the programs with different longer-term impacts differ.

Declining Effects Over Time? These patterns raise the question of whether program effect sizes have changed systematically across cohorts. Unfortunately, we know of no study with a credible design that has been able to test this question across a broad set of programs.

However, evidence from multiple studies converges on the conclusion that evaluations of early education programs are producing smaller effects now than what was reported for previous cohorts.

First, a systematic examination of early childhood program impacts between 1960 and 2007 found that the average end-of-treatment impacts on cognitive outcomes were smaller for recent programs compared with older programs (Duncan & Magnuson, 2013). Second, recent quasi-experimental studies support this conclusion as well. While Deming's (2009) sibling fixed-effect Head Start study found generally positive impacts into early adulthood, a replication study using more recent cohorts (Pages et al., 2020) found null and even negative impacts on school-based and longer-run outcomes such as cognitive tests, teen parenthood, educational attainment, and health for Head Start attendees.

The most recent evidence on the impacts of early childhood programs within and across cohorts comes from the Meta-Analysis of Educational RCTs with Follow-up (MERF), a dataset comprised of post-test and follow-up impacts on cognitive and social-emotional outcomes from a broad range of educational RCTs systematically sampled from eight existing meta-analyses (for more details see Hart et al., 2024). In the supplementary information, we provide more details regarding this meta-analytic dataset, including information for obtaining a public-use version of the data.

Figure 1 depicts the average effects of 18 treatment-control contrasts, all from RCTs or lottery studies, targeting children prior to elementary school entry outside of the home (see also supplemental Table S2). The figure's bold lines reflect the meta-analytic averages across all interventions, whereas the fainter lines display the average posttest and follow-up effects for each treatment-control contrast. The figure shows considerable variability in the impacts for programs across time, especially for studies with small sample sizes (e.g., the large 6- to 12-month follow-up effect for the Abecedarian study). Despite this variation, the average end-of-treatment impact for programs that started in 1960-1999 (.47 SD) is more than twice as large as that for programs that started in 2000 to 2011 (.21 SD; $p_{difference}$ = .14). When considering effects averaged across all assessment waves together, older programs produced larger impacts (.24 SD vs. .08 SD; $p_{difference}$ = .04) than more recent programs. Moreover, even conditional on post-test effect sizes, the impacts of the more recent programs appear to fade out more quickly than the impacts for the older evaluations (e.g., 61% vs. 39% of post-test impact at first follow up).

One might be concerned that these cohort differences simply reflect changes in the focus of interventions (i.e., programmatic vs. curricular) and outcomes researchers have assessed (i.e., cognitive vs. social-emotional) overtime. While the data suggest that there have certainly been changes across time in both the focus of the intervention and the outcomes measured in their evaluations, these changes do not appear to fully account for the differences in initial impacts and persistence (see supplemental Table S2 and Figures S1-S4). In particular, for *programmatic* interventions (i.e., attending preschool), impacts were both substantially smaller and faded more quickly as a proportion of the end of treatment impact in more recent evaluations.

Possible Explanations for Declining Preschool Impacts

All estimates of preschool program impacts involve comparisons of outcomes for children who were offered or otherwise received program services with children who were not offered or did not receive those services. Explanations of changing impacts over time must therefore consider both changes in program effectiveness and changes in the experiences and environments of children not offered or participating in the program.

Improved Counterfactual Conditions. Societal changes in the last 50 to 60 years have improved the home and other environmental conditions for most children and families. This likely reduces the value-added of some of the relative advantages provided by early preschool programs in the 1960s and 70s. The most noteworthy environmental enrichments are in 1) the quality of children's home environments, partly because of dramatic increases in supports available for low-income families; 2) the expansion of community-based preschool (e.g., mixed delivery preschool, for-profit centers), which have expanded parents' options for early care and education outside of public preschool programs. We discuss each domain below.

Secular Trends in Children's Environments. In the 1960s and 1970s, children from low-income families, especially Black children, faced deplorable conditions. In the mid-1960s, the Food Stamp program and programs such as the Earned Income Tax Credit had not yet been introduced. Racial discrimination in parts of the country denied Black children access to quality schools and Black families access to hospital care, including physician-assisted childbirth (Reynolds, 2004). Many safety-net programs for children in low-income families were introduced and others markedly expanded between 1960 and 2000. As shown in Table 1, spending on welfare programs and ECE services has risen dramatically, with a large increase seen in Medicaid and cash assistance (AFDC and SNAP) funding both prior to and after 2000, as well as substantial increases in Head Start and Child Care and Development Fund dollars.

Home environments have improved as well. Mothers with low incomes completed 2 more years of schooling in 2000 compared with 1960, and family size shrunk by over 30%. Additionally, parents spend more time with their children now: in the 1990s, parents increased the amount of time they spent with their children dramatically, by nearly a third for non-college-educated mothers and nearly 100% for non-college-educated fathers (Ramey & Ramey, 2009).

Expenditures on child "enrichment goods" increased sharply between the early 1970s and early 2000s for both low and higher-income families (Kornrich & Furstenberg, 2013).

Such broad environmental improvements have likely improved developmental outcomes for most low-income children and reduced the value added by public ECE programs. For example, child health has improved dramatically in the U.S., with infant and child mortality falling by over 70% between 1960 and 2000 (See Table 1). Racial and income-based achievement gaps have shrunk over time, particularly for the children with the lowest academic proficiency (Latham, 2018; Reardon & Portilla, 2016), an improvement that might be caused, at least in part, by increased ECE access. Not all environmental changes were positive, as single motherhood and crime rose over the same period. Still, the changes to family size, maternal education, child health, ECE participation, and public expenditures on children's health, nutrition, and education, suggest that families with low incomes today face more favorable conditions for raising their children compared with families with low incomes raising children in the middle of the 20th century.

Even assuming constant program quality across time, low-quality environmental conditions made it much easier for initial Head Start programs or a model program like Perry Preschool to demonstrate effectiveness for enrolled children when compared with children experiencing business-as-usual conditions. Consistent with this explanation, the positive long-run effects of Head Start, even in earlier cohorts, were smaller when children had more access to Food Stamps and when they lived in counties with lower poverty levels (Bailey et al., 2021).

Greater Availability of Preschool Services. Another explanation for the mixed effects is the greater availability of center-based child care for children ages five and under, especially among low-income children typically targeted by public preschool programs (Cascio, 2021; see

Table 1). This would reduce the contrast between children who were and were not offered pre-k or Head Start program slots (i.e., substitution bias). Kline and Walters (2016) found that the experimental impact of Head Start on early academic skills was larger and more likely to be sustained in the short run when the children who attended Head Start would otherwise have stayed home with their parents or been cared for in a home-based preschool setting. Yet public programs are not and could not be offered only to children who otherwise would stay home.

Preschool Program Quality and Content. Another potential explanation for the mixed findings in this literature is variability in program quality within modern-day preschool programs (Bloom & Weiland, 2015; Sabol et al., 2020). As we note above, access to state pre-k and Head Start programs expanded dramatically, and this rapid scale up most certainly involved variable implementation quality. The quality of Head Start's classroom processes today is substantially higher than the early years of the program because the performance quality standards became more rigorous over time (U.S. DHHS, 2023). In contrast, state pre-k programs vary substantially with respect to program standards, funding, and monitoring (Friedman-Krauss, et al., 2023).

But this variability in program standards, and thus in the quality of preschool classroom experiences, does not explain why most pre-k evaluations show moderate to large impacts initially that diminish in the school years (Phillips et al., 2017). In Tennessee, quality was sufficiently high enough to boost treatment children's initial skills compared with children in other care settings (Lipsey et al., 2018); Tennessee also met 9 of the 10 quality benchmarks used by preschool researchers (Friedman-Krauss et al., 2023). The high levels of structural quality and lack of process quality data make it impossible to explain why Tennessee's positive impacts turn negative in middle school (Durkin et al., 2022), or why programs with similar large initial impacts differ in effectiveness during elementary school. However, changes to specific elements of preschool programs—which are part of overall program quality—may contribute to declining or differing effects, which we discuss below.

Changes in Instructional Modes and Foci Over Time. Abecedarian and Perry focused on strong caregiver-child relationships with frequent multi-turn conversations, and hands-on learning activities in which teachers scaffolded learning (e.g., Ramey et al., 2014). Head Start initially focused on promoting health–a crucial need in the 1960s (Zigler & Styfco, 2010). In 2007, federal guidance began requiring Head Start programs to teach early literacy and math skills to address kindergarten-entry gaps between low- and middle-income children (U.S. DHHS, 2007). Similarly, in 2009, the National Association for the Education of Young Children (NAEYC) updated their position statement, highlighting achievement gaps and emphasizing the need to meaningfully align preschool and elementary school standards and content, a change from the original 1986 and revised 1996 position statements (NAEYC, 1986; 2009).

Practice guidance and research foci are often intertwined, driving one another forward based on the current perspectives (Peisner-Feinberg & Yazejian, 2010). Accordingly, how this guidance translated into practice is difficult to discern given limited longitudinal data on program instructional practices, time spent on specific content areas, and whether instruction was in centers, individual work, and/or small or large groups. A comparison of nationally representative Head Start program data from 2001 to 2015, based on teacher-report, showed children spending a growing proportion of time in language and literacy activities, particularly from 2001 to 2007, and an increase in teacher-directed activities, particularly in full-day classrooms (Markowitz & Ansari, 2020). Thus, this change from using hands-on learning and focusing on promoting health, language, and social skills to using teacher-directed instruction to teach early academic skills could contribute to changes in the longer-term findings across time.

Public preschool programs are funded by state and local governments and vary widely in how they are structured and implemented, precluding generalizations about the instructional focus and content of this collection of programs. Data from recent different preschool evaluations offer some evidence regarding the amount of time spent in literacy and math instruction, along with more whole-group instruction (e.g., Burchinal et al., 2021; Denker & Attenbery, 2024; Justice et al., 2022). For example, in one study, researchers found preschoolers in a Colorado public school district in half-day programs in 2017-2019 were spending 51 minutes or 27% of their time in whole class instruction (Denker & Atteberry, 2024). The number of minutes in academic content areas per day varied widely by teacher (e.g., range of approximately 18 to 39 minutes for English Language Arts). Burchinal and colleagues (2021) found similar patterns, with preschool children spending 36% of their time in large group settings, and 39% in a literacy or math activity in North Carolina pre-k classrooms in 2016-2017. Most of this academic instruction occurred in large groups and was didactic teacher-led instruction. Not all whole group instruction is didactic and of course sometimes can be engaging for preschoolers, but the observed amount of time spent in this format is well above the guidance of using large group time for brief instructions or explanations (Burchinal et al., 2022).

As to declining program effectiveness over time, it could be the case that hands-on experientially oriented learning of the earlier preschool programs added something unique to children's schooling experiences, and that is perhaps no longer true today. Although Abecedarian lacked an explicit focus on teaching academic skills, Abecedarian children entered kindergarten with large advantages in language and IQ and smaller advantages in reading and math skills over their control-group counterparts. Program impacts on both reading and math skills persisted or grew in second grade and were maintained through 14-21 years of age (Campbell et al., 2012;

García et al., 2020). In contrast, the strongest causal evaluations of contemporary programs that focus on teaching academic skills reported impacts on K-12 academic skills that are either negative or null (Durkin et al., 2022; Gray-Lobe et al., 2023; Puma et al., 2012).

The basic skills focus of current programs may optimize for impacts in the pre-k year (Phillips et al., 2017), but many children in the control group may make up much of the kindergarten-entry skill gaps by the end of the year because these same skills are often taught again in kindergarten (Cohen-Vogel et al., 2021). Indeed, as preschool programs shifted toward a curricular focus, kindergarten instruction also changed from focusing on socializing young children to teaching early reading and math skills (Bassok et al., 2016). Accordingly, all groups of children enter kindergarten with higher average levels of academic skills compared with previous cohorts (Bassok & Latham, 2017; Latham, 2018) and kindergarten teachers often focus on teaching these same skills (Engel et al., 2013). Thus, instructional redundancy in pre-k and kindergarten (Cohen-Vogel et al., 2021) may contribute to the convergence of skill levels over time for preschool attenders and nonattenders.

Curriculum Evaluations. Because RCT evaluations of preschool programs estimate impacts of the entire preschool package, which is comprised of curricula, teachers, peers, and setting, they cannot answer whether one curriculum promotes children's short and long-term achievement better than another. RCT evaluations of preschool curricula are needed to assess whether focused content does have lasting effects on children's outcomes. Clear evidence from rigorous studies indicates that targeted academic curricula boosts children's school readiness skills at the end of the preschool year (NASEM, 2024; Jenkins et al., 2018). Yet, follow-up studies show that curricular impacts often fade or disappear completely during elementary school (Bailey et al., 2018; Preschool Curriculum Evaluation Research Consortium, 2008), casting doubt on the proposition that implementing curricula targeted on children's achievement or educational outcomes will generate long-run impacts. Although short-term evidence that targeted curricula improve school readiness skills is reassuring, evidence that those short-term advantages are maintained after children transition to elementary school is needed to support arguments that high-quality curriculum and aligned coaching will best promote children's development (Weiland et al., 2018; NASEM, 2024).

Other Widely Discussed Fadeout Explanations

Scaling Up Research Projects. In hindsight, it is naïve to believe that findings from the early RCT studies would generalize to public programs operating at scale that serve millions of children each year at a fraction of the per-child cost. Operating an at-scale program almost always involves lower costs for serving each child to make the program financially feasible. Launching at-scale programs also entails receiving buy-in from the many stakeholders and almost always requires changing the program to meet their goals and needs (Tseng et al., 2017). While these factors clearly explain part of why researcher-run programs have larger impacts, they are less likely to explain reduced impacts on adulthood outcomes in Head Start (Deming, 2009; Pages et al., 2020).

Quality of Subsequent Schooling. A widely accepted explanation for fading impacts is that the quality of K-12 schools attended by children from preschool programs serving low-income families are too low to build on the skills children acquired in preschool (Abenavoli, 2019). One compelling analysis of the effects of Head Start attendance has found larger long-run impacts on adult outcomes for children who attended elementary schools that received additional funding during the 1970s and 80s (Johnson & Jackson, 2019). However, a comprehensive meta-analysis of studies testing the sustaining environment hypothesis found little evidence to support this explanation (Bailey et al., 2020). This explanation is further complicated by evidence from a recent evaluation that suggests attending a higher quality elementary school can accelerate fadeout by speeding up the time when children without preschool experience catch up with children attending preschool (Watts et al., 2023).

Conclusion

Many factors likely explain why today's preschool programs less consistently impart the kinds of long-term academic advantages that earlier demonstration programs did. Research shows the promise of some modern-day programs (Gray-Lobe et al., 2023), yet the mechanisms through which programs are effective across the lifecourse is unknown. Changes in the scale of preschool programs, who administers the programs, what they target, increased opportunities and safety net coverage for low-income minoritized families likely account for why today's programs have smaller impacts at entry to kindergarten. Those arguments, however, cannot account for impacts that fade in most longer-term evaluations and, especially, why initially positive impacts turned negative in one randomized evaluation. Further worrisome is that these high-quality programs are often the ones supported by public funds serving children from low-income families with the goal of promoting equity.

It is unreasonable to expect that results from expensive early demonstration programs would carry over to today's scaled-up and lower cost programs. We argue instead that researchers, policymakers, and practitioners must adjust expectations on what today's programs can produce in terms of promoting children's development in the long run. Although some scholars may argue that mixed results from preschool evaluations are to be expected and are good for policymaking, if we cannot predict a priori the long-term effects based on program components, then such heterogeneity is likely to be idiosyncratic and unhelpful to policymakers.

We argue the mixed findings from modern day preschool programs require researchers to update theories and investigate mechanisms to help understand how programs can best promote children's short and long-term development. Further, preschool serves multiple goals, including providing a safe, reliable care setting for children and promoting positive development and well-being (Burchinal et al., 2022). Our paper focuses on the latter, but the former is an equally important goal. More research on the long-term benefits of preschool access for parents and families should be conducted.

We make two recommendations regarding future research to further understand modern-day long-term preschool effectiveness. First, additional long-term follow-up studies are needed from RCTs so that we understand the connections between short- and longer-term effects. The long-term follow-up studies of children who attended pre-k over 20 years provided surprising evidence of young adult impacts, even in the absence of short or intermediate impacts (Gray-Lobe et al., 2023). Careful follow-up studies of rigorous RCTs like the Head Start Impact Study or pre-k lottery studies are needed to determine whether they too will show these young adult impacts despite the fading of the initial short-term impacts. Such evaluations could allow for systematic investigation of the extent to which long-term impacts are forecasted by short-term impacts, with major implications for developmental theory and intervention design.

Second, research should examine whether changes in environmental conditions and in program focus and style of instruction explain, at least in part, why today's programs' impacts vary from positive, albeit smaller, null, or negative, and fade faster when children enter elementary school. The almost complete fadeout of today's preschool programs in RCT studies, and negative longer-term results in at least one evaluation, suggest that research must describe the classroom experiences of children in each program to identify instructional practices and content that best promote long-term success. Longitudinal research to identify trends in preschool time use and instructional formats is needed and will help in further understanding effective preschool components. Specifically, future research can examine whether different patterns of longer-term impacts could be obtained with the instructional practices recommended in the National Academies' *A New Vision for High Quality Pre-K Curriculum* Report (NASEM, 2024) and used in the earlier programs–which are similar to those recommended in a recent What Works Clearinghouse guide (Burchinal et al., 2022) on preparing young children for school (i.e., language-rich, hands-on learning activities, and frequent conversations between teachers and students).

We believe the field should take seriously the null and negative results arising from recent preschool program evaluations, and, rather than dismiss the findings using explanations that do not hold when thoroughly examined, strive to understand how best to make preschool programs effective in the current conditions.

References

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

Amadon, S., Gormley, W. T., Claessens, A., Magnuson, K., Hummel-Price, D., & Romm, K. (2022). Does early childhood education help to improve high school outcomes? Results from Tulsa. *Child Development, 93*(4), e379-e395. https://doi.org/10.1111/cdev.13752

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

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

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, M. J., Sun, S., & Timpe, B. (2021). Prep School for poor kids: The long-run impacts of Head Start on Human capital and economic self-sufficiency. *American Economic Review, 111*(12), 3963-4001. https://doi.org/10.1257/aer.20181801

Barnett, W. S., & Jung, K. (2021). Effects of New Jersey's Abbott preschool program on children's achievement, grade retention, and special education through tenth grade. *Early Childhood Research Quarterly, 56*, 248-259. https://doi.org/10.1016/j.ecresq.2021.04.001

Bassok, D., & Latham, S. (2017). Kids today: The rise in children's academic skills at kindergarten entry. *Educational Researcher, 46*(1), 7-20. https://doi.org/10.3102/0013189x17694161

Bassok, D., Latham, S., & Rorem, A. (2016). Is kindergarten the new first grade? *AERA Open, 2*(1). https://doi.org/10.1177/2332858415616358

Bloom, H. S., & Weiland, C. (2015). Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study. *Available at SSRN.* https://doi.org/10.2139/ssrn.2594430

Burchinal, M., Garber, K., Foster, T., Bratsch-Hines, M., Franco, X., & Peisner-Feinberg, E. (2021). Relating early care and education quality to preschool outcomes: The same or different models for different outcomes? *Early Childhood Research Quarterly, 55*, 35-51. https://doi.org/10.1016/j.ecresq.2020.10.005

Burchinal, M., Krowka, S., Newman-Gonchar, R., Jayanthi, M., Gersten, R., Wavell, S., Lyskawa, J., Haymond, K., Bierman, K., Gonzalez, J. E., McClelland, M. M., Nelson, K., Pentimonti, J., Purpura, D. J., Sachs, J., Sarama, J., Schlesinger-Devlin, E., Washington, J., & Rosen, E. (2022). Preparing Young Children for School (WWC 2022009). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. Retrieved from https://whatworks.ed.gov/.

Burchinal, M., Whitaker, A., Jenkins, J., Bailey, D., Watts, T., Duncan, G., & Hart, E. (2024). Unsettled science on longer-run effects of early education. *Science, 384*(6695), 506-508. https://doi.org/10.1126/science.adn2141

Burchinal, M., Whitaker, A. A., & Jenkins, J. M. (2022). The promise and purpose of early care and education. *Child Development Perspectives, 16*(3), 134-140. https://doi.org/10.1111/cdep.12463

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), 1033-1043. https://doi.org/10.1037/a0026644

Cascio, E. U. (2021). *Early Childhood Education in the United States: What, When, Where, Who, How, and Why*. NBER Working Paper #28722. https://doi.org/10.3386/w28722

Cohen-Vogel, L., Little, M., Jang, W., Burchinal, M., & Bratsch-Hines, M. (2021). A missed opportunity? Instructional content redundancy in Pre-K and kindergarten. *AERA Open, 7*. https://doi.org/10.1177/23328584211006163

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

Denker, H., & Atteberry, A. (2024). Where has all the time gone? Describing time use in full-vs. half-day pre-Kindergarten. *Early Childhood Research Quarterly, 68*, 235-246. https://doi.org/10.1016/j.ecresq.2024.05.007

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

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

Elango, S., García, J. L., Heckman, J. J., & Hojman, A. (2016). Early childhood education. In *Economics of means-tested transfer programs in the United States, volume 2* (pp. 235-297). University of Chicago Press. https://doi.org/10.7208/chicago/9780226392523.003.0004

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching Students What They Already Know? The (Mis)Alignment Between Mathematics Instructional Content and Student Knowledge in Kindergarten. *Educational Evaluation and Policy Analysis, 35*(2), 157-178. https://doi.org/10.3102/0162373712461850

Friedman-Krauss, A., Barnett, W. S., Hodges, K. S., Garver, K., Weisenfeld, G. G., Gardiner, B. A., & Jost, T. M. (2023). *The State of Preschool 2022*. https://nieer.org/wp-content/uploads/2023/05/YB2022_FullReport.pdf

Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of Head Start. *American Economic Review, 92*(4), 999-1012. https://doi.org/10.1257/00028280260344560

García, J. L., Bennhoff, F. H., Leaf, D. E., & Heckman, J. J. (2021). *The dynastic benefits of early childhood education*. NBER Working Paper #315555. https://doi.org/10.3386/w31555

García, J. L., Heckman, J. J., Leaf, D. E., & Prados, M. J. (2020). Quantifying the life-cycle benefits of an influential early-childhood program. *Journal of Political Economy, 128*(7), 2502-2541. https://doi.org/10.1086/705718

García, J. L., Heckman, J. J., & Ronda, V. (2023). The lasting effects of early-childhood education on promoting the skills and social mobility of disadvantaged African Americans and their children. *Journal of Political Economy, 131*(6), 1477-1506. https://doi.org/10.1086/722936

Gormley, W. T., Phillips, D., & Gayer, T. (2008). Preschool programs can boost school readiness. *Science, 320*(5884), 1723-1724. https://doi.org/10.1126/science.1156019

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

Hart, E. R., Bailey, D. H., Luo, S., & Watts, T. W. (2024). *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). https://doi.org/10.26300/7j8s-dy98

Jenkins, J. M., Duncan, G. J., Auger, A., Bitler, M., Domina, T., & Burchinal, M. (2018). Boosting school readiness: Should preschool teachers target skills or the whole child? *Economics of Education Review, 65*, 107-125. https://doi.org/10.1016/j.econedurev.2018.05.001

Johnson, R. C., & Jackson, C. K. (2019). Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending. *American Economic Journal: Economic Policy, 11*(4), 310-349. https://doi.org/10.1257/pol.20180510

Kline, P., & Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of Head Start. *The Quarterly Journal of Economics, 131*(4), 1795-1848. https://doi.org/10.1093/qje/qjw027

Kornrich, S., & Furstenberg, F. (2013). Investing in children: Changes in parental spending on children, 1972–2007. *Demography, 50*(1), 1-23. https://doi.org/10.1007/s13524-012-0146-4

Latham, S. (2018). Changes in school readiness of America's entering kindergarteners (1998–2010). In A. Mashburn, J. LoCasale-Crouch, & K. Pears (Eds.), *Kindergarten transition and readiness: Promoting cognitive, social-emotional, and self-regulatory development*, pg. 111-138. Springer Cham. https://doi.org/10.1007/978-3-319-90200-5_5

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

Logan, J. A., Piasta, S. B., Purtell, K. M., Nichols, R., & Schachter, R. E. (2024). Early childhood language gains, kindergarten readiness, and Grade 3 reading achievement. *Child Development, 95*(2), 609-624. https://doi.org/10.1111/cdev.14019

Markowitz, A. J., & Ansari, A. (2020). Changes in academic instructional experiences in Head Start classrooms from 2001–2015. *Early Childhood Research Quarterly, 53*, 534-550. https://doi.org/10.1016/j.ecresq.2020.06.008

NAEYC. (1986). NAEYC Position Statements on Developmentally Appropriate Practice in Early Childhood Programs Serving Children From Birth Through Age 8. *Young Children, 41*(6), 3-19. http://www.jstor.org/stable/42725817

NAEYC. (2009). NAEYC Position Statements on Developmentally Appropriate Practice in Early Childhood Programs Serving Children From Birth Through Age 8 - Position Statement. https://www.naeyc.org/sites/default/files/globally-shared/downloads/PDFs/resources/position-statements/PSDAP.pdf

National Academies of Sciences, Engineering, and Medicine. 2024. A new vision for high-quality preschool curriculum. Washington, DC: The National Academies Press. https://doi.org/10.17226/27429.

Pages, R., Lukes, D. J., Bailey, D. H., & Duncan, G. J. (2020). Elusive longer-run impacts of head start: replications within and across cohorts. *Educational Evaluation and Policy Analysis, 42*(4), 471-492. https://doi.org/10.3102/0162373720948884

Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., & Weiland, C. (2017). Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects. *The current state of scientific knowledge on pre-kindergarten effects*, 19-30.

Preschool Curriculum Evaluation Research Consortium. (2008). Effects of preschool curriculum programs on school readiness (NCER 2008-2009). *Washington, DC: National Center for Education Research, Institute of Education Sciences, US Department of Education. Washington, DC: US Government Printing Office*.

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, OPRE Report # 2012-45, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Ramey, G., & Ramey, V. A. (2009). *The rug rat race*. NBER Working Paper #15284 https://doi.org/10.3386/w15284

Ramey, C. T., Sparling, J. J., & Ramey, S. L. (2014). Interventions for students from low resource environments: The Abecedarian approach. In J. Mascolo, V. Alfonson, & D. Flanagan (Eds.). *Essentials of planning, selecting, and tailoring interventions for unique learners*, 415-448. John Wiley & Sons.

Reardon, S. F., & Portilla, X. A. (2016). Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *AERA Open, 2*(3), https://doi.org/10.1177/2332858416657343

Reynolds, P. P. (2004). Professional and hospital discrimination and the US Court of Appeals Fourth Circuit 1956–1967. *American Journal of Public Health, 94*(5), 710-720. https://doi.org/10.2105/ajph.94.5.710

Sabol, T. J., Ross, E. C., & Frost, A. (2020). Are all Head Start classrooms created equal? Variation in classroom quality within Head Start centers and implications for accountability systems. *American Educational Research Journal, 57*(2), 504-534. https://doi.org/10.3102/0002831219858920

Tseng, V., Easton, J. Q., & Supplee, L. H. (2017). Practice Partnerships: Building Two-Way Streets of Engagement. Social Policy Report. Volume 30, Number 4. *Society for Research in Child Development*. https://doi.org/10.1002/j.2379-3988.2017.tb00089.x

U. S. Department of Health and Human Services (2007). Head Start Act. Retrieved from: https://eclkc.ohs.acf.hhs.gov/sites/default/files/pdf/HS_Act_2007.pdf

U. S. Department of Health and Human Services (2023). History of Head Start. Retrieved from: https://www.acf.hhs.gov/ohs/about/history-head-start.

Watts, T. W., Jenkins, J. M., Dodge, K. A., Carr, R. C., Sauval, M., Bai, Y., Escueta, M., Duer, J., Ladd, H., & Muschkin, C. (2023). Understanding Heterogeneity in the Impact of Public Preschool Programs. *Monographs of the Society for Research in Child Development, 88*(1), 7-182. https://doi.org/10.1111/mono.12463

Weiland, C., McCormick, M., Mattera, S., Maier, M., & Morris, P. (2018). Preschool curricula and professional development features for getting to high-quality implementation at scale: A comparative review across five trials. *AERA Open, 4*(1) https://doi.org/10.1177/2332858418757735

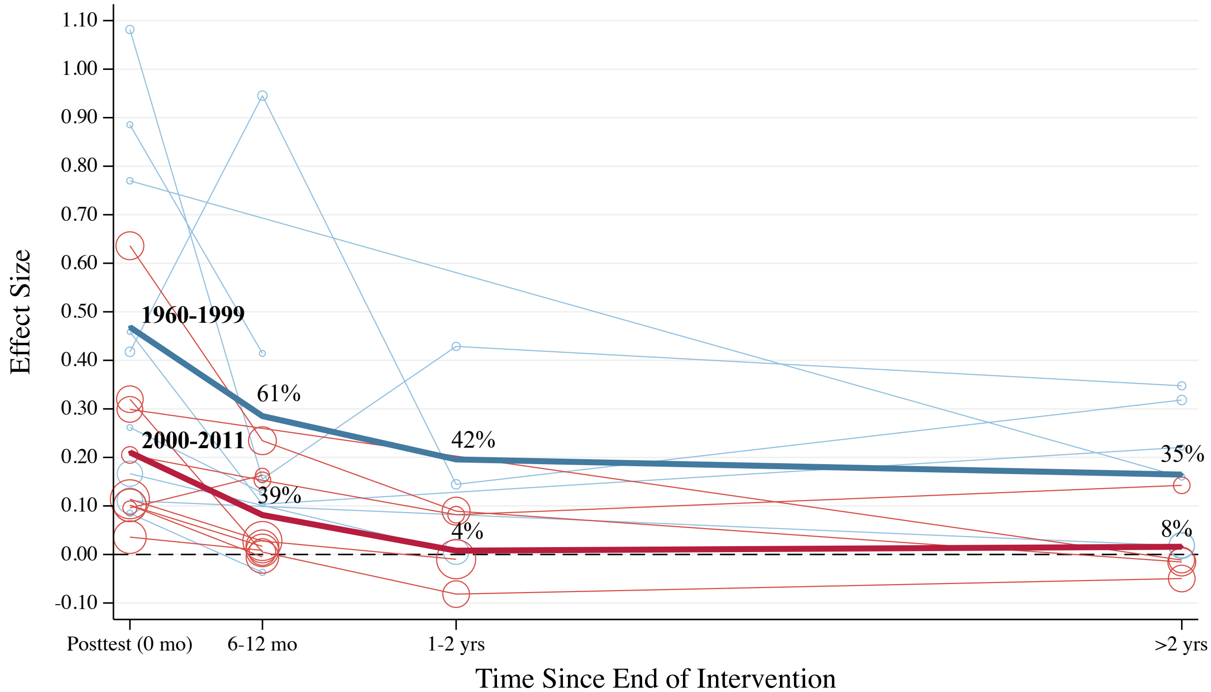
Weiland, C., Unterman, R., Shapiro, A., Staszak, S., Rochester, S., & Martin, E. (2020). The effects of enrolling in oversubscribed prekindergarten programs through third grade. *Child Development, 91*(5), 1401-1422. https://doi.org/10.1111/cdev.13308

White House. (2023). *Economic Report of the President 2023.* https://www.whitehouse.gov/wp-content/uploads/2023/03/ERP-2023.pdf

Zigler, E., & Styfco, S. J. (2010). *The hidden history of Head Start*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195393767.001.0001

Figure 1.

Average Treatment Impacts for Selected Sample of Early Childhood Educational Interventions by Implementation Timing



Note: Blue = interventions that began in 1960-1999 (n = 9 treatment groups). Red = interventions that began in 2000-2011 (n = 9 treatment groups). Interventions targeting children outside of the home prior to elementary school entry were included in this selected sample of early childhood educational programs from the broader MERF sample (see Hart et al., 2024). The bold lines represent the meta-analytic average of all available effects at each assessment wave from models that included weighting by 1/SE^2 and a study random effect. Each coordinate represents the average intervention effect for an intervention treatment group formed through randomization. All social-emotional and cognitive effects reported at a given time point contributed to the averages. Lines connect the average effects across assessment waves within a particular treatment group. Coordinates are weighted by 1/SE^2 where SE is the average post-test SE for the respective treatment group. The x-axis ticks are scaled by the average time elapsed since treatment end for each bin (e.g., the average time since post-test for a greater than 2-year follow-up assessment was 94 months. The x-axis is scaled accordingly). Percentages reflect the percentage of the average post-test effect observed at each follow-up wave (respectively for 1960-1999 and 2000-2011 interventions).

Table 1.
Conditions in the United States, Expenditures, Child Health, Crime, and ECE Participation Across Time

	1960		**2000**		**2019**	
	Bottom Income Quintile	**National Average**	**Bottom Income Quintile**	**National Average**	**Bottom Income Quintile**	**National Average**
Demographics for families with children <5						
Number of Children in Home	3.8	3.8	2.8	3.2	2.5	2.8
Maternal Labor Force Participation	21.9%	20.6%	57.4%	66.8%	55.2%	64.6%
Maternal Education Level (in years)	9.1	10.7	11.3	12.6	11.3	12.8
Single Mothers	3.4%	0.8%	34.4%	14.4%	37.4%	14.4%
Federal Expenditures (in billions, 2023$)	**1960**		**2000**		**2019**	
Medicaid	0		40.6		117	
AFDC/TANF	6		19.8		15.1	
SNAP	0		16.7		32.8	
WIC	0		6		5.6	
Head Start	0		7.6		11.2	
Child Care Development Fund	0		5.7		8.6	
Per Capita Expenditures (2023$)	**1960**		**2000**		**2019**	
Total Federal Child Expenditures (per capita)	362		3,407		6,230	
Per Pupil Public K-12 Spending	4,535		14,211		17,910	
Child Health	**1960**		**2000**		**2019**	
Infant (< 1 year) Mortality (per 100k)	2,700		728.7		558.3	
Child (1- 4 years) Mortality (per 100k)	110		32.9		23.3	
Crime	**1960**		**2000**		**2019**	
Violent Crime (per 100k)	160.9		506.5		366.7	
Property Crime (per 100k)	1,726.3		3,618.3		2,109.9	
ECE Participation (% Enrolled)	**1970**		**2000**		**2019**	
Age 3 Child Care or Pre-K	12.9%		39.2%		53.8%[a]	
Age 4 Child Care or Pre-K	27.8%		64.9%			
Age 5 Child Care, Pre-K, or Kindergarten	69.3%		87.6%		90.7%	

Note. Demographic information comes from authors' calculations of the harmonized Decennial Census and American Community Surveys made available by the Integrated Multiuse Public Microdata Series (IPUMS) at the University of Minnesota (Ruggles et al., 2023).

Expenditure figures are in billions of dollars and are in 2023 dollars. Total child expenditures is per child capita. Federal expenditure and federal child per capita expenditure information comes from https://www.urban.org/sites/default/files/publication/102614/kids-share-2020-chartbook_0.pdf. Per pupil K-12 spending information comes from https://nces.ed.gov/programs/digest/d22/tables/dt22_236.55.asp. Specifically, we use data from expenditure per pupil in fall enrollment, total expenditures. 1959-60; 1999-00; 2018-19 (National Center for Education Statistics [NCES], 2023a).

Infant mortality rate is number of infant deaths before age 1 per 100,000. Child mortality rate is number of child deaths (ages 1-4) per 100,000. Data from 1960 comes from: https://www.cdc.gov/nchs/data/vsus/VSUS_1960_2A.pdf; Data from 2000 comes from: https://www.cdc.gov/nchs/data/nvsr/nvsr50/nvsr50_15.pdf; Data from 2019 comes from: https://www.cdc.gov/nchs/data/databriefs/db395-H.pdf.

Crime data comes from the U.S. Census Bureau for 1960 and 2000 (https://www2.census.gov/library/publications/2004/compendia/statab/123ed/hist/hs-23.pdf). The 2019 data comes from Federal Bureau of Investigation (https://www.fbi.gov/news/press-releases/fbi-releases-2019-crime-statistics#:~:text=The%202019%20statistics%20show%20the,2%2C109.9%20offenses%20per%20100%2C000%20inhabitants.). Violent crime includes murder, manslaughter, rape, robbery, and aggravated assault. Property crimes include burglary, larceny-theft, and motor vehicle theft.

Early childhood education participation information comes from NCES (NCES, 2023b): https://nces.ed.gov/programs/digest/d19/tables/dt19_202.10.asp, and https://nces.ed.gov/programs/digest/d22/tables/dt22_202.20.asp. Specifically, we used percent of enrolled 3, 4, and 5 year olds in school.
a = In 2019, the number of 3 and 4 year olds enrolled in school was not available individually for each age, thus the combined percentage of 3 and 4 year olds enrolled in school is displayed.

Supplemental Materials for:
"Are Preschool Programs Becoming Less Effective?"

Meta-Analysis of Educational RCTs with Follow-up

In this manuscript, we used data from the larger Meta-Analysis of Educational RCTs with Follow-up (MERF) dataset to investigate the descriptive differences in posttest and follow-up impacts of early childhood educational programs. Hart et al., (2024) provides in-depth information about the creation of the MERF sample. The data and project protocols are available for review and use via LDbase: https://doi.org/10.33009/ldbase.1719529626.152e. In brief, 305 studies from eight existing meta-analyses were screened for inclusion in MERF. 85 studies (110 treatment-control contrast groups) met inclusion criteria as educational RCTs that reported impacts on cognitive and/or social-emotional outcomes, measured follow-up impacts at least six months after intervention end, and provided usable statistics that allowed effect size estimation. To identify whether studies met these criteria, an extensive search process was conducted to gather all available effect sizes reported across assessment waves. The included studies were diverse in treatment focus and in what developmental period was targeted; the sample was comprised of preschool interventions, elementary-level curricular interventions, substance use prevention programs, etc.

For the purposes of this manuscript, we restricted the sample to studies that met the following criteria: 1) targeted children prior to elementary school entry; 2) a major component of the intervention occurred outside of the home environment (i.e., programs that solely provided home visiting would be excluded); 3) reported impacts on at least one cognitive (achievement, general cognition, language or literacy, math) or social-emotional (internalizing, externalizing, general social-emotional skills) outcome at post-test or follow-up. With these criteria, we were able to include 14 studies that contained 18 treatment-control contrasts (i.e., some studies had multiple treatment groups that were formed through randomization). 8 studies (9 treatment-control contrasts) were initiated in 1960-1999 and 6 studies (9 treatment-control contrasts) were initiated in 2000-2011. A list of the studies that we included is provided in Table S1 below. We collapsed the data into the following assessment wave bins: post-test, 6- to 12-month follow-up, greater than 24-month follow-up. In cases that there was more than one assessment for a particular construct, measure, subscale, and reporter during each binned assessment window (e.g., 6-month and 9-month follow-up), we calculated the average effect size and standard error to produce one effect for analysis. We were able to include 363 effect sizes in total. Effects were estimated at the treatment-control contrast level.

To meta-analyze the data, we used the metafor package in R. For each assessment wave, we meta-analyzed all of the available outcomes with weighting by $1/se^2$ to upweight more precise estimates from treatment-control contrasts with larger samples. Because most studies/treatment-control contrasts contributed multiple outcomes, we included a study-level random effect in our models. While our main interest was in posttest and follow-up effect averages based on intervention year, we also considered differences by intervention focus, and outcome type. For intervention focus, we categorized the interventions as either being programmatic (i.e., introduced the treatment child to a *new* context that the control children did not get) or curricular (i.e., changed the treatment children's experience of an existing program/context). For outcome type, we were interested in the distinction between social-emotional outcomes (e.g. measures categorized as measuring general social-emotional skills, or internalizing, or externalizing skills more specifically) and cognitive outcomes (e.g., general cognitive skills such as intelligence, or, more specifically, measures of achievement, math, or language/literacy).

To test the statistical significance of differences in the effects of older and more recent programs, we ran a series of tests. First, we set out to identify whether there were end-of-treatment differences in the meta-analytic averages for programs that were older and more recent. We first ran two separate meta-analytic models to test for average end-of-treatment impacts for older programs and more recent programs with weighting by $1/se^2$ and study-level random effects. To test for whether this difference in means was statistically significant, we ran one model in which all posttest assessments were included form older and newer programs and intervention timing was entered as a predictor. This model contained the same features as the other models, with a random effect for study and weighting by $1/se^2$. The *p*-value associated with the intervention timing predictor was interpreted as the statistical significance of the difference in end-of-treatment effects by intervention timing.

Second, to test for the differences across all assessment waves, we first ran two models in which all effects from older programs and all effects from more recent programs were included, respectively. We used weighting by $1/se^2$ and a study random effect. Additionally, we entered a random effect for “aligned group” to account for the fact that several outcomes were measured consistently across posttest and follow-up assessment waves and, thus, were represented in the model multiple times. Aligned groupings were cases where the same construct was assessed using the same measure, subscale, and reporter at more than one assessment wave within the same study and treatment-control contrast group. There were many cases in which an “aligned group” contained $n = 1$ outcomes, reflecting that the outcome was only collected at one assessment wave. Finally, to test for the statistical significance of the difference in these means, we ran a model in which effects from all assessments (posttest and follow-up) from older and more recent assessments were included. Intervention timing was entered as a predictor as were three dummy variable covariates capturing assessment timing (with posttest excluded as the reference group). This model similarly used weighting by $1/se^2$ and random effects for study and aligned group. The *p*-value associated with the intervention timing predictor was interpreted as the statistical significance of the difference in all treatment effects by intervention timing.

Table S1.
Treatment-Control Contrasts Included in the Sample by Intervention Timing

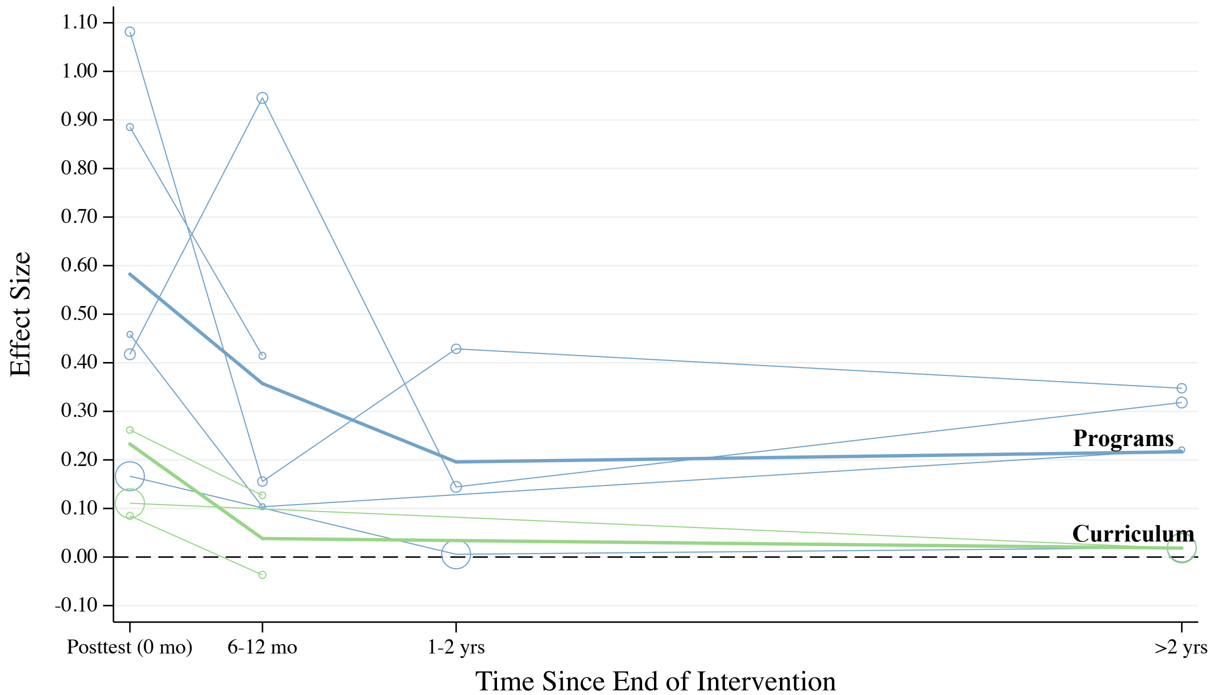
1960-1999
Abecedarian~ All Pre-K (tx + tx; tx + cntrl) Group
Classroom and At-Home Preschool Interventions
Dialogic Reading #1~ School plus Home Reading Group
Dialogic Reading #1~ School Reading Group
Early Head Start
Infant Health and Development Program
Perry Preschool
The Early Training Project~ 3-year and 2-year Intervention Group
Project Know How
2000-2011
Head Start Impact Study
Head Start Research-Based, Developmentally Informed Program
Head Start CARES~ Incredible Years Teacher Training Group
Head Start CARES~ Preschool Promoting Alternative Thinking Strategies Group
Head Start CARES~ Tools of the Mind Play Group
Teacher Responsivity Education
TRIAD~ Building Blocks Group
TRIAD~ Building Blocks plus TRIAD Follow-Through Group
Tennessee Pre-K

Table S2.
Average Intervention Impacts Across Decades for Early Childhood Educational Interventions

	All Outcomes				Cognitive Outcomes				Social-Emotional Outcomes				Programmatic Interventions				Curricular Interventions			
Assessment Wave	ES (*SE*)	*p*	n_o	n_c	ES (*SE*)	*p*	n_o	n_c	ES (*SE*)	*p*	n_o	n_c	ES (*SE*)	*p*	n_o	n_c	ES (*SE*)	*p*	n_o	n_c
All Intervention Years																				
Treatment End	0.34 (0.08)	0.00	134	18	0.35 (0.08)	0.00	83	18	0.09 (0.03)	0.00	51	7	0.46 (0.14)	0.00	35	7	0.21 (0.07)	0.00	99	10
6 mo to 1 yr	0.12 (0.04)	0.00	75	14	0.13 (0.05)	0.01	42	14	0.07 (0.05)	0.15	33	6	0.21 (0.12)	0.07	32	6	0.11 (0.05)	0.02	43	8
> 1 yrs, up to 2 yrs	0.08 (0.06)	0.21	50	7	0.07 (0.06)	0.25	22	6	0.09 (0.11)	0.41	28	4	0.08 (0.10)	0.38	44	5	0.08 (0.04)	0.03	6	2
> 2 yrs	0.10 (0.05)	0.05	104	10	0.11 (0.08)	0.15	47	9	0.05 (0.03)	0.17	57	6	0.15 (0.09)	0.11	59	5	0.06 (0.04)	0.18	45	4
1960-1999																				
Treatment End	0.47 (0.14)	0.00	38	9	0.48 (0.14)	0.00	32	9	0.05 (0.02)	0.00	6	2	0.58 (0.18)	0.00	20	5	0.23 (0.15)	0.11	18	3
6 mo to 1 yr	0.29 (0.15)	0.05	26	6	0.32 (0.14)	0.03	20	6	0.01 (0.08)	0.92	6	1	0.36 (0.18)	0.05	20	4	0.04 (0.15)	0.79	6	2
> 1 yrs, up to 2 yrs	0.20 (0.14)	0.18	16	3	0.18 (0.14)	0.19	10	3	0.46 (0.10)	0.00	6	1	0.2 (0.14)	0.18	16	3	--	--	--	--
> 2 yrs	0.16 (0.07)	0.02	69	6	0.18 (0.09)	0.03	39	6	0.03 (0.01)	0.04	30	4	0.22 (0.09)	0.02	47	4	0.02 (0.01)	0.11	22	1
2000-2011																				
Treatment End	0.21 (0.06)	0.00	96	9	0.21 (0.07)	0.00	51	9	0.11 (0.05)	0.01	45	5	0.21 (0.10)	0.05	15	2	0.21 (0.09)	0.02	81	7
6 mo to 1 yr	0.08 (0.04)	0.03	49	8	0.06 (0.03)	0.09	22	8	0.08 (0.06)	0.17	27	5	0.03 (0.01)	0.07	12	2	0.12 (0.05)	0.02	37	6
> 1 yrs, up to 2 yrs	0.01 (0.04)	0.84	34	4	-0.01 (0.01)	0.45	12	3	-0.01 (0.05)	0.78	22	3	-0.04 (0.04)	0.24	28	2	0.08 (0.04)	0.03	6	2
> 2 yrs	0.02 (0.07)	0.81	35	4	-0.08 (0.05)	0.13	8	3	0.06 (0.08)	0.41	27	2	-0.08 (0.01)	0.00	12	1	0.07 (0.08)	0.35	23	3

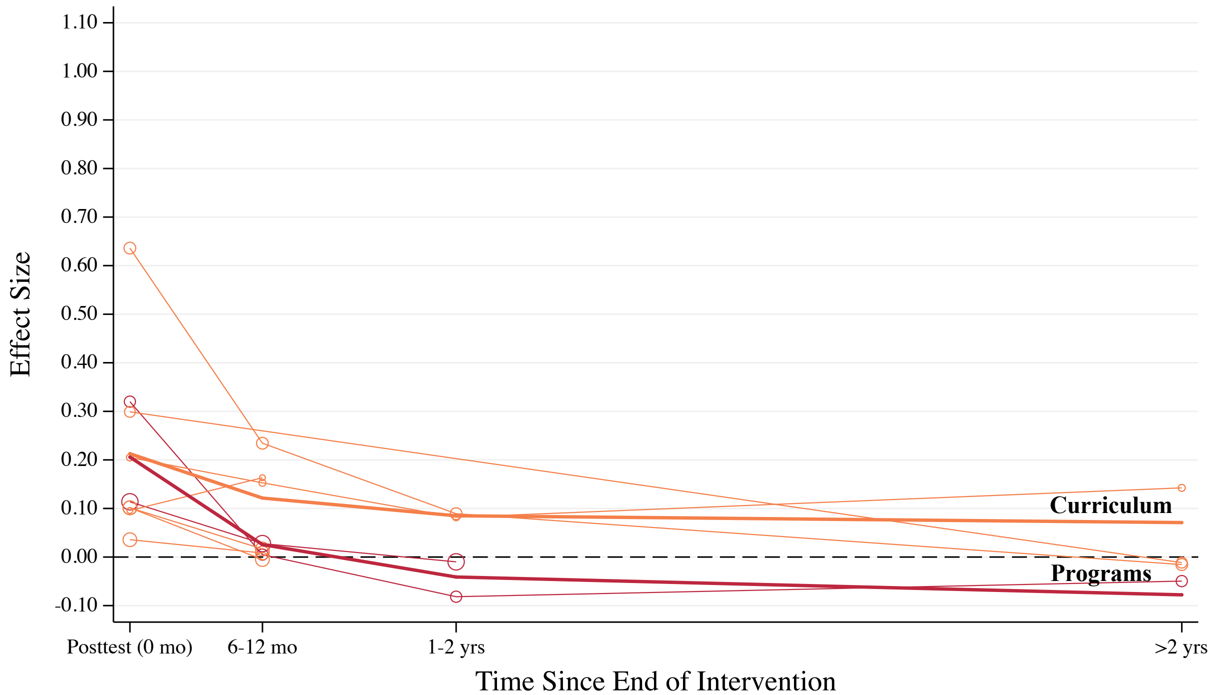
Note: This table presents meta-analytic average treatment impacts from models that included weighting by 1/SE^2 and a study-level random effect. Only early childhood educational interventions that targeted children outside of the home prior to elementary school entry were included in these models from the broader MERF sample (see Hart et al., 2024). Outcomes that contributed to the social-emotional outcomes were measures categorized as those that tapping general social-emotional skills, or internalizing, or externalizing skills specifically. Outcomes that contributed to the cognitive outcomes included measures of general cognitive skills such as intelligence, or, more specifically, measures of achievement, math, or language/literacy. Programmatic interventions involved the receipt of a new context (e.g., pre-k program for those in the treatment group, no pre-k program for those in the control group), whereas curricular interventions involved a change to a child's existing context (e.g., new reading curriculum for those in the treatment group). ES = effect size in standard deviation units. SE = standard error. n_o = number of outcomes. n_c = number of treatment-control contrasts.

Figure S1
Average Treatment Impacts for 1960-1999 Interventions, by Intervention Type



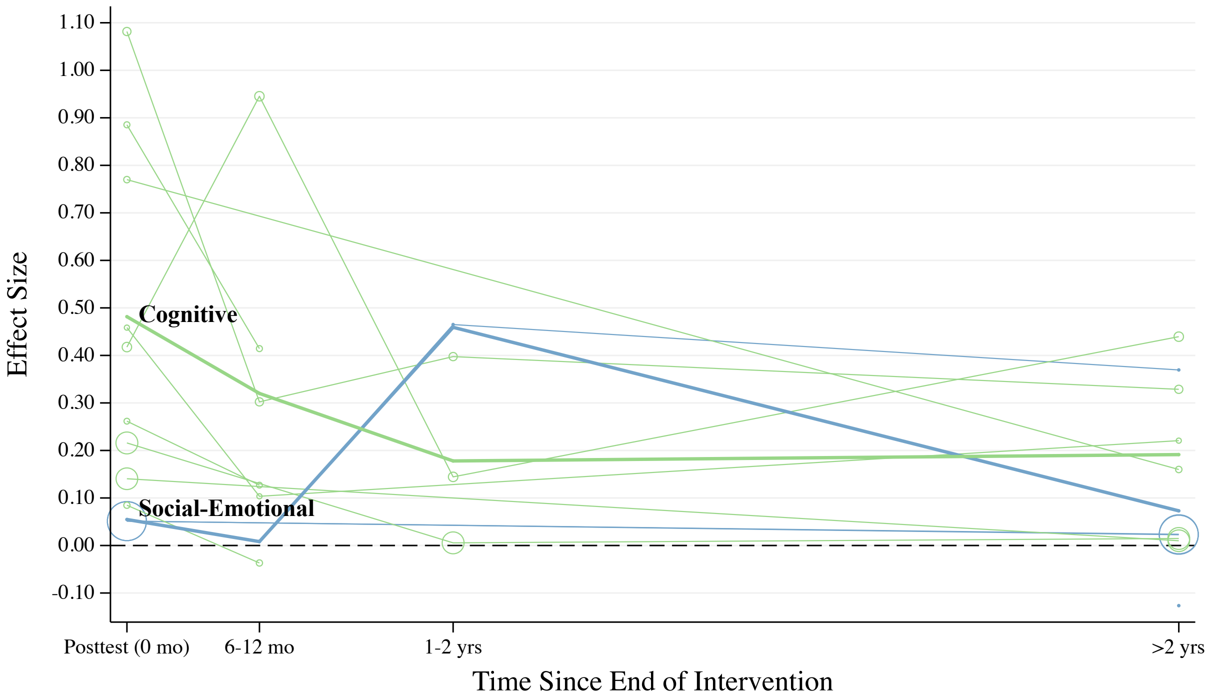
Note: Blue= programs (n = 5 treatment groups). Green = curricular interventions (n = 4 treatment groups). Only early childhood educational interventions that started between 1960 and 1999 and targeted children outside of the home prior to elementary school entry were included in this figure from the broader MERF sample (see Hart et al., 2024). The bold lines represent the meta-analytic average of all available effects at each assessment wave from models that included weighting by 1/SE^2 and a study random effect. Each coordinate represents the average intervention effect for an intervention treatment group formed through randomization. All social-emotional and cognitive effects reported at a given time point contributed to the averages. Lines connect the average effects across assessment waves within a particular treatment group. Coordinates are weighted by 1/SE^2 where SE is the average post-test SE for the respective treatment group. The x-axis ticks are scaled by the average time elapsed since treatment end for each bin. For consistency, the average time elapsed from all early childhood educational interventions included in Figure 1 was used.

Figure S2
Average Treatment Impacts for 2000-2011 Interventions, by Intervention Type



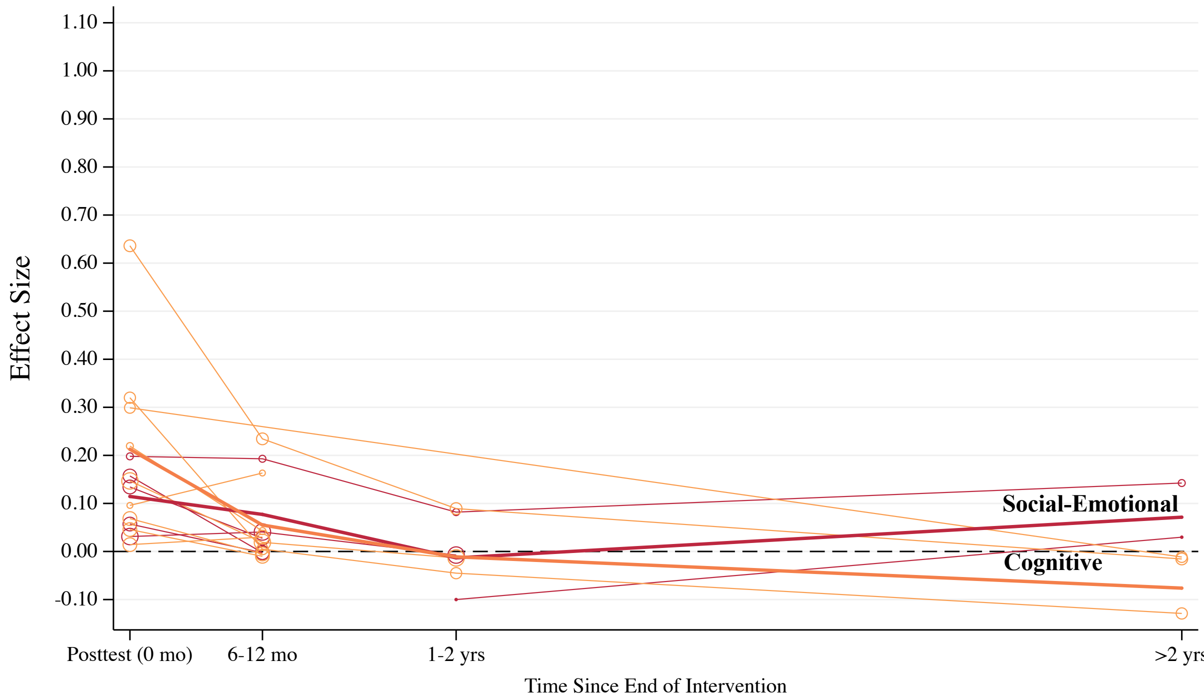
Note: Dark red = programs (n = 2 treatment groups). Orange = curricular interventions (n = 7 treatment groups). Only early childhood educational interventions that started between 2000 and 2011 and targeted children outside of the home prior to elementary school entry were included in this figure from the broader MERF sample (see Hart et al., 2024). The bold lines represent the meta-analytic average of all available effects at each assessment wave from models that included weighting by 1/SE^2 and a study random effect. Each coordinate represents the average intervention effect for an intervention treatment group formed through randomization. All social-emotional and cognitive effects reported at a given time point contributed to the averages. Lines connect the average effects across assessment waves within a particular treatment group. Coordinates are weighted by 1/SE^2 where SE is the average post-test SE for the respective treatment group. The x-axis ticks are scaled by the average time elapsed since treatment end for each bin. For consistency, the average time elapsed from all early childhood educational interventions included in Figure 1 was used.

Figure S3
Average Treatment Impacts for 1960-1999 Interventions, by Outcome Type



Note: Blue = social-emotional (n = 4 treatment groups). Green = cognitive (n = 9 treatment groups). Only early childhood educational interventions that started between 1960 and 1999 and targeted children outside of the home prior to elementary school entry were included in this figure from the broader MERF sample (see Hart et al., 2024). The bold lines represent the meta-analytic average of all available effects at each assessment wave from models that included weighting by 1/SE^2 and a study random effect. Each coordinate represents the average intervention effect for an intervention treatment group formed through randomization. All social-emotional and cognitive effects reported at a given time point contributed to the averages. Lines connect the average effects across assessment waves within a particular treatment group. Coordinates are weighted by 1/SE^2 where SE is the average post-test SE for the respective treatment group. The x-axis ticks are scaled by the average time elapsed since treatment end for each bin. For consistency, the average time elapsed from all early childhood educational interventions included in Figure 1 was used. Of note, there are not four lines for the social-emotional outcomes, as suggested by n=4 treatment groups, because two of the treatment groups contributed social-emotional outcomes to the meta-analytic averages at at least one assessment wave, but did not measure any social-emotional outcomes at post-test.

Figure S4
Average Treatment Impacts for 2000-2011 Interventions, by Outcome Type



Note: Dark red = social-emotional (n = 6 treatment groups). Orange = cognitive (n = 9 treatment groups). Only early childhood educational interventions that started between 2000 and 2011 and targeted children outside of the home prior to elementary school entry were included in this figure from the broader MERF sample (see Hart et al., 2024). The bold lines represent the meta-analytic average of all available effects at each assessment wave from models that included weighting by 1/SE^2 and a study random effect. Each coordinate represents the average intervention effect for an intervention treatment group formed through randomization. All social-emotional and cognitive effects reported at a given time point contributed to the averages. Lines connect the average effects across assessment waves within a particular treatment group. Coordinates are weighted by 1/SE^2 where SE is the average post-test SE for the respective treatment group. The x-axis ticks are scaled by the average time elapsed since treatment end for each bin. For consistency, the average time elapsed from all early childhood educational interventions included in Figure 1 was used. Of note, there are not six lines for the social-emotional outcomes, as suggested by n=6 treatment groups, because one of the treatment groups contributed social-emotional outcomes to the meta-analytic averages at least one assessment wave, but did not measure any social-emotional outcomes at post-test.