Planning and Changing
Vol. 51, No. 3/4, 2024, pp. 74–87

# CONSIDER ALL THE EVIDENCE ON PER-K PROGRAMS FOR LOW-INCOME CHILDREN: WHY RANDOMIZED CONTROLLED TRIAL RESULTS MUST NOT DICTATE PUBLIC POLICY

*The recently published randomized controlled trial (RCT) study of the Tennessee Voluntary Pre-Kindergarten (VPK) program by Durkin et al. (2022) has sparked wide interest among educators, researchers, and policymakers. The study's finding that sixth graders who participated in VPK in school years 2009–10 and 2010–11 performed more poorly on their academic and behavioral outcomes than children who did not participate has been used to argue against public investment in pre-K programming. We join other education researchers and advocates in urging caution in interpreting the results and particularly in generalizing them to apply to today's pre-K programs in Tennessee or to programs in other states or cities. Our concerns are related to the challenges of implementing RCTs and using RCTs in studying early childhood education programming. Using the Tennessee VPK study as an example, we illustrate how those challenges can hinder the generalization of the findings. We also do not agree that the Tennessee VPK findings should—at least, not in isolation—lead to an assumption that public pre-K cannot improve outcomes for disadvantaged children. Evidence from other kinds of studies beyond RCTs should also inform decisions about public investment in pre-K education. Equally important is the need to "open the black box" to find out what is actually happening in pre-K classrooms—the approaches and interventions that are most effective, as well as what can happen after pre-K that helps sustain early benefits.*

Keywords: randomized controlled trial (RCT), public pre-K, early childhood policy

## Introduction

What is at stake in the ongoing debate over publicly funded pre-K programs is whether such an investment truly helps bridge the opportunity gap between more and less affluent children and families. The latest evidence in that ongoing debate is the most recent findings of the randomized controlled trial (RCT) of the Tennessee Voluntary Pre-Kindergarten (VPK) program (Durkin et al., 2022). The study found that sixth graders who participated in VPK in school years 2009–10 and 2010–11 performed more poorly on standardized tests and had more disciplinary problems





than children who did not participate (Durkin et al., 2022). Some commentators (e.g., Goodkind, 2022; Levitz, 2022) have used such findings to argue against public investment in pre-K programming. In response, some education researchers and advocates have argued for caution in interpreting the results and particularly in generalizing them to apply to today's pre-K programs in Tennessee or to programs in other states or cities (Barnett, 2022; Weiland et al., 2022).

We join other education researchers in urging caution in the use of the Tennessee and other pre-K RCT results for pre-K policy decisions. Calls to defund public pre-K, in particular, are based on a misunderstanding of the import of RCT findings; cutting access to a vital service for disadvantaged families is likely to reinforce the structural racism and socioeconomic oppression that put them at risk in the first place.

The authors of the Tennessee study, who are highly respected researchers in the field, have implemented a solid research design whose findings command respect. However, we cannot agree that their findings can or should be generalized to apply to other publicly funded pre-K programs, past or present. We also cannot agree that these findings should—at least, not in isolation—lead to an assumption that public pre-K puts disadvantaged children further at risk in the long run. Our reasons have to do with the limitations of RCT designs in education research and specifically in pre-K interventions; this article systematically outlines some of those issues. Furthermore, the few RCTs in the field constitute a tiny portion of a vast literature on the short-, medium-, and long-term effects of pre-K programming. Evidence from other kinds of studies beyond RCTs should also inform decisions about public investment in pre-K education. Equally important is the need to "open the black box" to find out what pre-K approaches and interventions are most effective. In addition, researchers and policymakers should attend to the well-established link between the sustainability of pre-K's early benefits and the quality of children's subsequent education.

## The Challenges of Using RCTs in Early Education

RCTs are generally accepted as a scientific method to evaluate cause-and-effect relationships; some consider RCTs to be the "gold standard" for deriving causal inferences (Grimes, 1991; Shadish et al., 2002; Sullivan, 2011). The most appealing feature of RCTs is that, when properly implemented, they create treatment and control groups that are statistically balanced on known and unknown confounding factors. With this balance in place, researchers can attribute observed differences between the treatment and control groups to the treatment rather than to any preexisting differences between the groups (Shadish et al., 2002). This feature also eliminates selection bias in treatment assignment and yields unbiased estimates of causal effects that align with statistical sampling theories and

methods.

The RCT method was instituted in the agricultural sciences in the 1920s and in medicine in the 1940s (Armitage, 2003). The success of RCTs rests on the ability to create laboratory conditions in which confounding factors can be controlled (Morrison, 2001). In the social sciences in general, and educational programs in particular, such control is neither ethical nor feasible (Morrison, 2001; Sullivan, 2011). The challenges of using RCTs to study early childhood programming are illustrated by the fact that only two other published manuscripts to date, in addition to the Tennessee VPK study, have demonstrated the use of RCT designs for large-scale impact studies of public preschool programs: the Head Start impact study (Puma et al., 2012) and the more recent study of Boston's pre-K program (Weiland et al., 2020). We have grouped the challenges of implementing RCTs and using RCT results in studying early childhood education programming into four categories:

- The inability to blind participants
- Treatment contamination
  - Assignment noncompliance
  - Spillover effects
- The lack of representativeness of the sample
- The inability to control for post-randomization influences

## The Inability to Blind Participants

In early childhood education studies—unlike, for example, RCTs in medicine or agriculture—participants and families know whether or not they have been selected to receive the treatment. Lack of blinding in RCT can bias the results and weaken the validity of the inferences derived from the study, as medical (e.g., Karanicolas et al., 2010; Schulz et al., 1995) and medical education (Sullivan, 2011) researchers have pointed out. Grimes (1991) cautions that both random assignment and subsequent treatment (or lack of treatment) must be blinded, or the RCT can produce misleading results. Schulz et al. (1995), speaking of medical trials, go so far as to say that "without proper application of measures to achieve concealment, the whole point of randomization vanishes and bias is likely to distort results" (p. 412).

Yet blinding is virtually impossible in studies of educational interventions (Thomas, 2016). In pre-K RCTs, families know whether they have been assigned to the program, and it is unethical to prevent families in the control group from seeking alternative options for their children. Participants who know their treatment assignment status can modify their behavior in ways that influence subsequent outcome measures; for example, control group members may compensate for not receiving the treatment by working harder or getting external support, while treatment group members may relax their efforts simply because they have extra support that presumably benefits them (Conrad & Conrad, 2005). In the Boston RCT, the authors found that 97% of control group participants ended up attending some center-based preschool program. They noted that this level of center-based pre-K attendance is an "unusual counterfactual in the public pre-K evaluation literature" (Weiland et al., 2020, p. 1402). This finding might be a real example of compensation for comparison group assignment. The magnitude and impacts of such efforts may be hard to examine or quantify. Still, the possibility of this compensation limits the ability of the RCT to reveal the true effects of the treatment.

## Treatment Contamination

In pre-K RCTs, two types of treatment contamination are common: assignment noncompliance and spillover.

### *Assignment Noncompliance*

The change of roles between treatment and control group participants violates the assumption of assignment compliance, which is vital to the validity of RTC designs. In the Tennessee study (Durkin et al., 2022) and the other two large-scale early childhood RCT studies (Puma et al., 2012; Weiland et al., 2020), control group members changed their assignment status by attending the program when seats became available after random assignment. Meanwhile, some children assigned to the treatment group did not attend the program. Such role-switching, based on families' practical choices, is typical in educational program settings; however, assignment noncompliance can bias the results of an RCT (Keogh-Brown et al., 2007).

In some studies, researchers provide compensation to control group families in the form of guaranteed enrollment either in the same program in the next round or in other programs. For example, in the Head Start study, three-year-olds assigned to the control group were eligible to enroll in Head Start the next year as four-year-olds (Puma et al., 2012). Although such an approach is considered ethical, the effect of later enrollment or of enrollment in an alternative program dilutes the effect of the program being studied, making longitudinal comparisons of the effects on treatment and control groups difficult if not impossible.

To compensate for this common limitation of RCTs in social science, researchers often implement complier average causal effect (CACE) analyses (Keogh-Brown et al., 2007) to minimize the bias introduced by noncompliance. The Tennessee (Durkin et al., 2022) and Boston (Weiland et al., 2020) studies used this statistical method to compensate for cross-

over between treatment and control groups. However, the CACE method is effective in addressing contamination only if the contamination has been correctly documented (Keogh-Brown et al., 2007). The contamination produced by spillover, in which control group children benefit from interacting with treatment children, is difficult to measure and therefore cannot be addressed by CACE estimation.

*Spillover Effect*

As treatment and control children and families interact with each other over the many years between pre-K and third or sixth grade, control children may benefit from the treatment without directly receiving it. Early childhood educators and researchers are well aware of— and often welcome—such spillover effects. Some studies have shown that children who do not attend public pre-K programs but later attend schools or live in neighborhoods with high pre-K participation have better academic outcomes than children in schools or neighborhoods with lower rates of pre-K participation (Neidell & Waldfogel, 2010; Williams, 2019). Spillover is therefore a desirable effect for families and communities.

However, for RCTs of effective interventions, spillover reduces the gap between the treatment and control groups, leading to an underestimation of the treatment effect (Keogh-Brown et al., 2007; Williams, 2019). The larger and more pervasive a program is in a given community, the more likely it is that control and treatment families will interact, thereby producing spillover effects (List et al., 2019). Also, spillover that occurs over a prolonged period through school and neighborhood interactions is difficult if not impossible to track. One way to partially compensate for spillover and strengthen RCT estimates of the long-term effects of pre-K participation would be to control for the pre-K population of the child's cohort in school or community settings. However, this step was not taken in the Tennessee study or in any current preschool RCT literature.

The Lack of Representativeness of the Sample

The extent to which RCT findings can be applied to large populations depends on the representativeness of the sample and appropriate randomization of equally representative participants. Our concerns about the representativeness of the Tennessee VPK sample center around ways in which the oversubscribed sites on which the RCT design depends differed from other sites in Tennessee. In their report of results through grade 3 (Lipsey et al., 2018), the researchers identified differences between the oversubscribed sites and other program sites, including geographic concentration of oversubscribed sites in one region and over-concentration of partner sites as opposed to those run by school districts.

Furthermore, like many other state-funded pre-K programs (Friedman-Krauss, 2021), Tennessee VPK is designed primarily for low-income children but also admits children with other risk factors. In 2009–2011, when the study pre-K cohorts were defined, the other criteria included disability and English language learner status (Lipsey et al., 2013). If the Tennessee Department of Education followed then the procedure in place today (Tennessee Department of Education, 2020), low-income children were prioritized in admission, and then children with other risk factors were admitted if slots were still available. The implications of this eligibility and admission structure affect the representativeness of the sample. If applicants at an oversubscribed site consisted entirely of children who were eligible on the basis on income, then the treatment and control children were appropriately randomized but the site was not representative of all sites, because sites that were not oversubscribed were more likely to have room for children with secondary eligibility factors. If applicants at oversubscribed sites included children who were eligible on the basis of secondary factors, then randomly assigning all children, regardless of eligibility criteria, to be admitted or to be waitlisted would have utilized a different implementation policy because other programs prioritized income eligibility over other factors. In either case, the student populations in these oversubscribed sites can be expected to be different from those of sites that did not waitlist students.

The Tennessee researchers applied weighting factors to control for the observed characteristics of the sampled children (Lipsey et al., 2018). None of the study reports describe any attempt to control for differences among sites (Lipsey et al., 2013; Lipsey et al., 2018; Durkin et al, 2022). The random assignment of children to treatment or control conditions, irrespective of risk factors, is not common practice in state-funded pre-K programs. More generalizable findings might result from an RCT that reserves a percentage of slots for each stratum of children based on eligibility factors and randomizes within each stratum. Results could then be generated for the entire sample and for each subgroup.

The Inability to Control for Post-Randomization Influences

Good RCT studies present strong evidence when selection into control and treatment groups is completely random and the two groups are identical, so that the treatment is the only factor that can cause the effects. For the Tennessee and similar pre-K RCTs, control for subsequent influences on pre-K treatment and control group children and changes in their circumstances would inspire more confidence in the results. However, access to such follow-up data would require a level of data collection that may not be feasible in large-scale pre-K studies like the Tennessee RCT.

For example, an important data point to be included in the mod-

el is later school quality and teacher effectiveness. In a 2020 study based on the Tennessee RCT data, researchers connected the Tennessee VPK data with school performance data (Pearman et al., 2020). They found that VPK participants were most likely to maintain their academic advantage over nonparticipants when they experienced *both* high-quality schools *and* highly effective teachers after pre-K. They found no significant difference between treatment and control groups in the quality of their kindergarten teachers or schools (Pearman et al., 2020). Citing this equivalence, the Tennessee VPK researchers did not control for school quality or teacher effectiveness (Durkin et al., 2022). The problem is their assumption that school quality and teacher effectiveness remained unchanged from kindergarten, when Pearman et al. (2020) correlated VPK participation with school data, throughout elementary school and into grade 6 (Durkin et al., 2022). This assumption is problematic not only because children typically change teachers every year or often change school buildings between kindergarten and grade 6, but particularly because their previous study found that the ability of the VPK children to sustain their pre-K gains depended on *both* high-quality schools *and* highly effective teachers after pre-K (Pearman et al., 2020 p. 547). Failure to take into consideration a fundamental factor known to affect child outcomes poses a threat to external validity that should be acknowledged as a limitation.

## Considering All the Evidence

These concerns about RCT studies of pre-K interventions generally and the Tennessee study, in particular, suggest that policy and program decisions, when they affect children placed at risk, should not rely solely on RCT evidence. Decision-makers should also consider qualitative evidence from families and educators as well as quantitative evidence from quasi-experimental studies, including, for example, propensity score matching, difference in differences, and regression discontinuity designs. Many voices in education research have pointed out that the findings of RCTs have limited generalizability to settings beyond the ones studied. Meanwhile, ample evidence is available from other kinds of studies to add to the field's knowledge base. One problem is that nonacademic audiences—including policymakers—still tend to believe that the results of RCTs are "the truth" (Deaton & Cartwright, 2018). Our concern is that misusing RCT results to the point of cutting funding for public pre-K can have enormous consequences for disadvantaged children and families, reinforcing structural inequities by blocking access to a kind of intervention proven to promote economic and educational advancement (Bustamante et al., 2022). Ultimately, in order to inform policy and practice, the field needs a much better understanding not only of whether public pre-K programs are effective but particularly of what interventions most improve the outcomes of children from low-income backgrounds, how those interventions





work, and under what circumstances they are effective.

## The Limited Generalizability of RCTs

RCTs are generally accepted as the ideal means of establishing causal relationships. However, the careful conditioning necessary to design an RCT with strong internal validity often limits the external validity of its findings (Frieden, 2017)—that is, the extent to which the results can be applied in any situation beyond the one being studied. As Deaton and Cartwright (2018) put it, "Establishing causality does nothing in and of itself to guarantee that the causal relation will hold in some new case, let alone in general" (p. 12). They go on to say that even a perfectly designed RCT, one that is completely free of bias or confounding variables, would produce estimates of average treatment effects that apply only to the RCT sample, not to any other sample—even of participants in the same program at a different time or in a different setting (Deaton & Cartwright, 2018). This limitation alone should give pause to those who would use the Tennessee findings to argue that public pre-K in general does not work to improve outcomes for low-income children and therefore is not worthy of public investment.

The findings of the Tennessee study, to the extent that they achieve validity in light of the questions raised above, apply to the VPK cohorts of 2009–2010 and 2010–2011. Advocates have argued that program improvements and quality assurance systems implemented in TN-VPK since a quality improvement act in 2016 make today's program substantially different (Barnett, 2022; Tennesseans for Quality Early Education, 2022). Furthermore, the population of eligible families in Tennessee may have changed since 2009–2011. Nationwide, low-income families, on average, are better educated and have more access to early childhood programming than a decade ago (Bustamante et al., 2022; Phillips et al., 2017). For these and other reasons, a new study of the Tennessee VPK might well yield different results.

Finally, applying findings from one state's program as it was implemented more than a decade ago to all publicly funded pre-K programs today ignores the differences among those programs. As the latest State of Preschool report from the National Institute on Early Education Research shows, some programs target low-income families, while a few are universal. The mechanisms for enrolling eligible children differ. The report also outlines substantial differences in state policies governing teacher qualifications, classroom size, program content, quality assurance mechanisms, and a host of other factors known to influence educational quality (Friedman-Krauss et al., 2022). Generalizing from one state's program to all states' programs goes well beyond the level of evidence RCTs on education interventions can provide.

## The Need to Use Other Forms of Evidence

Meanwhile, although RCTs are valuable, they are not the only or even the most trustworthy source of information to guide policy decisions. As Thomas (2016) puts it, RCTs are one "part of the epistemological ecosystem of education inquiry" (p. 393). Designed to expose cause-effect relationships, they may not be capable of doing so in the complex contexts in which education takes place (Morrison, 2001; Norman, 2003; Thomas, 2016)—where caregivers, family members, teachers, program sites, schools, neighborhoods, media, and myriad other factors influence what happens to children.

In light of these considerations and the many challenges of implementing sound RCT designs in educational settings, researchers and policy makers should also consider the large body of evidence from careful quasi-experimental studies. As many researchers have pointed out (e.g., Sullivan, 2011), different research methods have different strengths and weaknesses. Compared to relying on a few RCTs, aggregating the findings of many diverse studies provides a more holistic picture of the landscape of public pre-K and the effectiveness of pre-K programs.

Meta-analyses of rigorous quasi-experimental studies have found, like the Tennessee RCT, that preschool helps make children ready for kindergarten (Burger, 2010; Camilli et al., 2010; Duncan & Magnuson, 2013; Yoshikawa et al. 2016). Effects on readiness skills have often been found to be more pronounced for children from economically disadvantaged backgrounds and for English language learners (e.g., Burger, 2010; Duncan & Magnuson, 2013; Phillips et al., 2017).

Less clear is how pre-K participation affects medium- and long-term outcomes. Some researchers have found that pre-K participation improves elementary school outcomes, particularly in cognitive domains (see, e.g., a meta-analysis by van Huizen & Plantenga, 2018). Many others have found, like the Tennessee and Head Start RCTs, that the positive effects of preschool fade out by grade 3 (Camilli et al., 2010; Duncan & Magnuson, 2013; Yoshikawa et al., 2016). The Tennessee RCT is, as the authors admit, the first to find negative effects in grade 6 (Durkin et al., 2022). Some longer-term quasi-experimental studies have found positive effects on academic and social outcomes in adolescence and youth adulthood (Burger, 2010; Duncan & Magnuson, 2013; McCoy et al., 2017; Vandell et al., 2010). The findings of many high-quality quasi-experimental studies on public pre-K should be given equal weight in policy decisions with the findings of the three RCTs.

## The Need to Discover What Works and What Doesn't

Ultimately, both policymakers and program leaders need data that generally are not yet available: findings on the mechanisms by which pre-K participation can affect later academic and social outcomes (Camilli et al., 2010; Heckman et al., 2013; Phillips et al., 2017). Research has established that quality matters: not only the quality of the preschool (e.g., Bustamante et al., 2022; Sylva et al., 2011; Vandell et al., 2010; Yoshikawa et al., 2016) but also the quality of later education (e.g., Bailey et al., 2017; Phillips et al., 2017; Yoshikawa et al., 2016)—as Pearman et al. (2020) found using data from the Tennessee study. We agree with Durkin et al. (2022) that more attention needs to be paid not only to *whether* pre-K programs work but *how* they work. Context matters; within a given program, implementation can vary widely, and individuals—site leaders, teachers and aides, children, caregivers—act independently (Morrison, 2001). As some of the best minds in pre-K evaluation have noted, the field needs to open the "black box" to discover what is happening at individual sites and in individual classrooms (Phillips et al., 2017, p. 2). In addition to findings from RCT and quasi-experimental quantitative studies, the field should add rich, context-sensitive data from qualitative studies of pre-K programming to learn about what works and what does not (Thomas, 2016).

Also necessary is careful attention to the interactions between pre-K education and the complex array of experiences that affect students' outcomes after they leave pre-K. In the long interval between pre-K and grade 6 or, better yet, between pre-K and young adulthood, what is the nature of children's educational experiences? How do their social environments affect their development? To concentrate solely on whether or not children participate in public pre-K is to explore only one mechanism among many that add up to a diverse set of effects. The kind of direct cause-effect relationship RCTs were designed to produce in medicine and agriculture is far simpler than what actually happens among children and families in their multiple contexts. The more and more varied kinds of data the field can amass, the better our policy decisions will be. In the meantime, policymakers should carefully consider all the currently available evidence in order to decide on funding for programs that benefit children from low-income backgrounds.

## Conclusion

In summary, we agree that the Tennessee study, like other RCT studies, provides important information to the pre-K literature. However, in light of the points discussed here, we stress the need to acknowledge the common limitations of RCT designs for generalizability and use for policy purposes. In most cases, the implementation of RCTs requires strict restrictions and conditions that qualify their external validity. Coupled with the state-to-state and back-then-and-now differences in pre-K programs, we recommend caution in the interpretation of this study's results beyond the Tennessee pre-K program that was in existence when the study was done. In addition, the study's findings on the positive effect at

kindergarten and the negative effect afterward suggest the need to evaluate other factors, such as subsequent school quality and teacher effectiveness, that might have interacted with children's pre-K experience to support, decrease or negate pre-K gains in later years. Considering these additional pieces of evidence would greatly enhance our understanding of pre-K impacts and lead to more robust and effective policy decision-making.

## References


Armitage, P. (2003). Fisher, Bradford Hill, and randomization. I*nternational Journal of Epidemiology*, 32(6), 925–928. https://doi.org/10.1093/ije/dyg286

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

Barnett, W. S. (2022, February 4). One swallow does not a summer make: Drawing valid inferences from the longitudinal evaluation of Tennessee pre-K outcomes. National Institute for Early Education Research. https://nieer.org/2022/02/04/one-swallow-does-not-a-summer-make-drawing-valid-inferences-from-the-longitudinal-evaluation-of-tennessee-pre-k-outcomes

Burger, K. (2010). How does early childhood care and educaiton affect cognitive development? An international review of the effects of early interventions for children from different social backgrounds. *Early Childhood Research Quarterly*, *25*, 140–165. https://doi.org/10.1016/j.ecresq.2009.11.001

Bustamante, A. S., Dearing, E., Zachrisson, H. D., & Vandell, D. L. (2022). Adult outcomes of sustained high-quality early child care and education: Do they vary by family income? *Child Development*, *93*, 502–523. https://doi.org/10.1111/cdev.13696

Camilli, G., Vargas, S., Ryan, S., & Barnett, W. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record*, *112*(3), 579–620. https://doi.org/10.1177/016146811011200303

Conrad, K. M., & Conrad, K. J. (2005). Compensatory rivalry. In B. Everitt & D. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (pp. 338–339). Wiley.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Duncan, G., & 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. http://dx.doi.org/10.1037/dev0001301

Frieden, T. R. (2017). Evidence for health decision making—beyond randomized, controlled trials. *New England Journal of Medicine*, *377*, 465–475. https://doi.org/10.1056/NEJMra1614394

Friedman-Krauss, A. H., Barnett, W. S., Garver, K. A., Hodges, K. S., Weisenfeld, G., Gardiner, B. A., & Jost, T. M. (2022). *The state of preschool 2021: State preschool yearbook*. National Institute for Early Education Research. https://nieer.org/state-preschool-yearbooks-yearbook2021

Goodkind, N. (2022, January 29). Democrats have wanted to spend billions on pre-K for years. But a new study reveals possible flaws with those programs. *Fortune*. https://fortune.com/2022/01/29/democrats-universal-prek-new-tennessee-study-negative-effects/

Grimes, D. A. (1991). Randomized controlled trials: "It ain't necessarily so". *Obstetrics & Gynecology*, *78*(4), 703–704.

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understand 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

Karanicolas, P. J., Farrokhyar, F., & Bhandari, M. (2010). Blinding: who, what, when, why, how? *Canadian Journal of Surgery*, *53*(5), 345.

Keogh-Brown, M. R., Bachmann, M. O., Shepstone, L., Hewitt, C., Howe, A., Ramsay, C. R., Song, F., Miles, J. N. V., Torgerson, D. J., Miles, S., Elbourne, D., Harvey, I., & Campbell, M. J. (2007). Contamination in trials of educational interventions. *Health Technology Assessment*, *11*(43), iii–107. https://doi.org/10.3310/hta11430

Levitz, E. (2022, February 5). Does pre-K actually hurt kids? *New York*. https://nymag.com/intelligencer/2022/02/does-pre-k-actually-hurt-kids.html

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

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013). *Evaluation of the Tennessee Voluntary Prekindergarten program: Kindergarten and first grade follow□up results from the randomized control design*. Peabody Research Institute, Vanderbilt University. https://my.vanderbilt.edu/tnprekevaluation/files/2013/10/August2013_PRI_Kand1stFollowup_TN-VPK_RCT_ProjectResults_FullReport1.pdf

List, J., Momeni, F., & Zenou, Y. (2019). *Are estimates of early education programs too pessimistic? Evidence from a large-scale field experiment that causally measures neighbor effects*. Centre for Economic Policy Research. https://cepr.org/active/publications/discussion_pa-

pers/dp.php?dpno=13725
McCoy, D., Yoshikawa, H., Ziol-Guest, K. M., Duncan, G. J., Schindler, H. S., Magnuson,, K., Yang, R., Koepp, A., & Shonkoff, J. P. (2017). Impacts of early childhood education on medium- and long-term educational outcomes. *Educational Researcher*, *46*(8), 474-487. https://doi.org/10.3102/0013189X17737739
Morrison, K. (2001). Randomised controlled trials for evidence-based education: Some problems in judging 'what works.' *Evaluation & Research in Education*, *15*(2), 69–83. https://doi.org/10.1080/09500790108666984
Neidell, M., & Waldfogel, J. (2010). Cognitive and noncognitive peer effects in early education. *Review of Economics and Statistics*, *92*, 562–576. https://doi.org/10.1162/REST_a_00012
Norman, G. (2003). RCT = results confounded and trivial: The perils of grand educational experiments. *Medical Education*, *37*, 582–584. https://doi.org/10.1046/j.1365-2923.2003.01586.x
Pearman, F. A., Springer, M. P., Lipsey, M., Lachowicz, M., Swain, W., & Farran, D. (2020). Teachers, schools, and pre-K effect persistence: An examination of the sustaining environment hypothesis. *Journal of Research on Educational Effectiveness*, *13*(4), 547–573. https://doi.org/10.1080/19345747.2020.1749740
Phillips, D., Lipsey, M., Dodge, K., Haskins, R., Bassok, D., Burchinal, M., Duncan, G. J., Dynarski, M., Magnuson, K. A., & Weiland, C. (2017). *Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects*. Brookings. https://www.brookings.edu/research/puzzling-it-out-the-current-state-of-scientific-knowledge-on-pre-kindergarten-effects/
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*. U.S. Department of Health and Human Services. https://www.acf.hhs.gov/sites/default/files/documents/opre/head_start_report_0.pdf
Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. (1995). Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*, 273(5), 408–412. https://doi.org/10.1001/jama.273.5.408
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth/Cengage Learning.
Sullivan, G. M. (2011). Getting off the "gold standard": Randomized controlled trials and education research. *Journal of Graduate Medical Education*, *3*(3), 285–289. https://doi.org/10.4300/JGME-D-11-00147.1
Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2011). Pre-school quality and educational outcomes at age 11: Low quality has little benefit. *Journal of Early Childhood Research*, *9*(2), 109–124. https://doi.org/10.1177/1476718X10387900
Tennessee Department of Education. (2020). Scope of services for voluntary pre-K 2020–21. https://www.tn.gov/content/dam/tn/education/health-&-safety/VPK_Scope%20of%20Services_2020_21.pdf
Tennesseans for Quality Early Education. (2022, February 2). Tennessee Voluntary Pre-K Study policy brief. https://tqee.org/newsroom/tennessee-voluntary-pre-k-study/
Thomas, G. (2016). After the gold rush: Questioning the "gold standard" and reappraising the status of experiment and randomized controlled trials in education. *Harvard Educational Review*, *86*(3), 390–411. https://doi.org/10.17763/1943-5045-86.3.390
Vandell, D., Belsky, J., Burchinal, M., Steinberg, L., Vandergrift, N., & NICHD Early Child Care Research Network. (2010). Do effects of early child care extend to age 15 years? Results from the NICH study of early child care and youth development. *Child Development*, *81*, 737–756. http://doi.org/10.1111/j.1467-8624.2010.01431.x
van Huizen, T., & Plantenga, J. (2018). Do children benefit from universal early childhood education and care? A meta-analysis of evidence from natural experiments. *Economics of Education Review*, *66*, 206–222. https://doi.org/10.1016/j.econedurev.2018.08.001
Weiland, C., Bassok, D., Phillips, D. A., Cascio, E. U., Gibbs, C., Stipek, D. (2022, February 10). What does the Tennessee pre-K study really tell us about public preschool programs? Brookings Institution. https://www.brookings.edu/blog/brown-center-chalkboard/2022/02/10/what-does-the-tennessee-pre-k-study-really-tell-us-about-public-preschool-programs/
Weiland, C., Unterman, R., Shapiro, A., Staszak, S., Rochester, S., & Martin, E. (2020). The effects of enrolling in oversubscribed pre-kindergarten programs through third grade. *Child Development*, *91*(5), 1401-1422. https://doi.org/10.1111/cdev.13308
Williams, B. J. (2019). The spillover benefits of expanding access to preschool. *Economics of Education Review*, *70*(1), 127-143. http://dx.doi.org/10.1016/j.econedurev.2019.04.002
Yoshikawa, H., Weiland, C., & Brooks-Gunn, J. (2016). When does preschool matter? *The Future of Children*, *26*(2), 21–35. https://doi.org/10.1353/foc.2016.0010

**Jamie Heng-Chieh Wu is an Associate Director for Community Evaluation Programs Office for Public Engagement and Scholarship University Outreach and Engagement and Research Assistant Professor at Michigan State University.**

**Hope Akaeze is a Project Statistician at Michigan State University.**