Aiming Further: Addressing the Need for High Quality Longitudinal Research in Education

By Tyler W. Watts*, Drew H. Bailey[+], and Chen Li*

Published in the *Journal of Research on Educational Evaluation*

On Dec 6, 2019:

https://doi.org/10.1080/19345747.2019.1644692

* Steinhardt School of Culture, Education and Human Development, New York University, 627 Broadway, 8th Floor, New York, NY, 10003 (e-mail: tyler.watts@nyu.edu).

[+] School of Education, University of California, Irvine, 3200 Education Drive, Irvine, CA 92697-5000

Word count: 3,994

Acknowledgement

The authors would like to thank Ana Auger, Greg Duncan, Dale Farran, Javanna Obregon, Cybele Raver, and Christina Weiland for their helpful comments on previous drafts. We would also like to acknowledge the support of the Institute for Educational Science (R305A160176) and the Jacobs Foundation. The content of this article is solely the responsibility of the authors and does not represent the views of those acknowledged, nor the views of the Institute of Educational Sciences or the Jacobs Foundation.

In educational research, the importance of longer-run follow-ups has been continually identified as a key priority for the field, with policy reports (Martin et al., 2018; McCormick, Hsueh, Weiland, & Bangser, 2017; Phillips et al., 2018), conference keynote addresses [see SREE invited lectures by Duncan (2015) and Singer (2019)], and "future directions" sections of research manuscripts noting the need to conduct evaluations with longitudinal follow-up. In recent years, the field has experienced substantial growth in the use of randomized control trials (RCTs) for the evaluation of educational programs, and at the same time, the wide availability of secondary administrative data sources has made longitudinal follow-up for these RCTs more possible than ever before (Penner & Dodge, 2019). However, despite these important innovations, educational interventions reporting long-run follow-up are still scarce, leaving a critical gap in the evaluation literature. In this commentary, we argue that this gap hampers the field's progress, stifling our ability to empirically test fundamental theories regarding long-run development, and incentivizing research practices that are counter-productive to our widely-held goals. Below, we offer several options that researchers and funders could pursue to substantially strengthen our understanding of how educational programs influence long-term student outcomes.

The Need for RCTs with Longitudinal Follow-Up

Educational research has benefitted greatly from longitudinal studies using correlational and quasi-experimental designs. Correlational studies have identified potential targets for educational interventions, and quasi-experimental studies have generated additional sources of data for estimating internally valid program impacts. However, quasi-experimental studies often carry limitations that complicate, or prevent altogether, longitudinal follow-up because the comparison group receives the treatment in a later period [e.g., age-of-entry regression

discontinuity designs for public pre-k (Weiland & Yoshikawa, 2013); difference-in-differences designs for school accountability (Dee & Jacob, 2011)]. Further, the instruments providing "exogenous" variation in most quasi-experimental studies are often subject to assumptions that are difficult to fully test, and correlational research is even further compromised by omitted variable bias. These limitations leave RCTs as the "gold standard" of evidence for educational program evaluation.[1]

Fortunately, educational RCTs have become much more common in recent years, partly due to the growing influence of The Institute for Education Sciences (IES). Since its inception in 2002, IES has become the dominant funder of educational intervention evaluations. Yet, despite explicitly calling for "follow-up" studies as part of its annual request for applications (RFA), a survey of funded IES projects highlights the severe lack of longitudinal follow-up in educational research. Using IES's public database of funded research grants and contracts (https://ies.ed.gov/funding/grantsearch/), we searched for all studies funded under the "Efficacy and Replication" category, which focuses on "the evaluation of fully-developed education interventions … [in] authentic education settings" and "follow-up studies of students" (i.e., "Goal 3" grants; see https://ies.ed.gov/funding/). This search returned 394 abstracts from funded grants, which we further narrowed to 370 abstracts that used some form of the term "random" and that we determined were RCTs. We then coded any study abstract that used the term "longitudinal," "follow-up," or "long-term" to record the furthest follow-up assessment planned post intervention (172 studies used one of these terms along with the term "random").

We found that only 27 of the 370 (7.3%) funded RCTs had discernable follow-up plans past 2 years after the end of the intervention. From this group of 27, 12 studies planned to follow students between 2 and 4 years post intervention, and 15 planned to follow students over 4 years.

The lack of longitudinal follow-up is partly due to the mechanics of IES funding, as grants do not typically last more than five years. However, we recorded only 20 studies that were dedicated follow-up studies with the purpose of tracking a sample that had already been examined in a previous evaluation. Thus, although the field has moved substantially toward RCT evaluations, these studies have largely lacked the longitudinal follow-up needed to assess whether the interventions in question make sustained impacts on child outcomes.

This research gap has led to several issues that continue to hamper the field's progress. First, researchers continue to rely on correlational evidence linking academic achievement measures to adult outcomes in order to project program impacts when follow-up measures are unavailable (Kraft, 2018). This projection is often made implicitly in introduction and discussion sections when researchers cite correlational studies to motivate the intervention at hand, or the projection is made explicitly when researchers use reported correlations between test scores and earnings to make labor-market impact projections for cost-benefit analyses (Krueger, 2003; Deming, 2009). Despite growing indications that this approach might provide inaccurate long-term estimates, both by under- (Bartik, 2014; Fredriksson, Öckert, & Oosterbeek, 2013) and over-estimating (Chetty et al., 2011) the long-run effect of interventions, it continues to be widely used (Bartik, Gormley, & Adelstein, 2012; Kline & Walters, 2016). This practice may lead researchers and practitioners to make inefficient investment decisions based solely on short-run impacts when long-term impacts are left unmeasured.

A single-minded focus on end-of-treatment outcomes also creates problematic incentives for researchers. As with the over-alignment problem between interventions and outcome measures (i.e., "teaching to the test"; see Slavin, 2008; Koretz, 2005), the alignment between the intervention and the *timing* of outcome measurement may incentivize curricula and pedagogy

tailored to a narrower set of academic skills than would be ideal for maximizing students' long-run success. This short-run focus may lead to interventions that are unlikely to complement students' subsequent educational experiences, while simultaneously creating little incentive for collaborative projects that would align children's educational experiences over a multi-year period (Stipek, Franke, Clements, Farran, & Coburn, 2017). The popular hypothesis that proposed interventions might be necessary, but not sufficient for spurring long-lasting change without complementary improvements to later educational quality, could be tested directly. Perhaps most importantly, focus on short-run impacts gives researchers few incentives to think about their own unique solutions to generating impacts on students' long-run outcomes, a problem that would benefit from diverse teams of researchers working toward the same goal (Brooks-Gunn, 2003).

Many factors likely contribute to the lack of longitudinal follow-up following educational interventions. Sample attrition following the end of treatment erodes study power over time, and researchers may consider several factors – the possibility of disappointing short-run fadeout on test scores; subsequent home, administrative, and curricular practices outside of the researcher's control; time that could be spent designing new interventions – as limiting the appeal of a resource-intensive follow-up. Indeed, substantial resources are required to collect follow-up data for large-scale RCTs, and as our coding exercise illustrated, researchers may simply lack the support needed to pursue follow-up studies. However, because we were not able to observe the complete pool of IES applications (only the studies that actually received funding), our coding exercise could not test whether the lack of follow-up funding was due to the applicant pool (i.e., few studies seek follow-up funding) or the grant selection process (i.e., follow-up studies are submitted but not selected). Given that the number of funded follow-up studies remains

remarkably low, it seems plausible that both the applicant pool and the grant selection process could benefit from a greater focus on longer-run outcomes. Consequently, we provide recommendations to funders and researchers in the sections below.

Recommendations for Funders

New RFAs should encourage researchers to pre-specify hypotheses for whether (and if so, how) their proposed intervention would affect long-term outcomes. For IES, this policy could ask researchers to incorporate their long-term hypotheses into their logic model, which would incentivize researchers to think carefully about the possible long-term implications of the intervention proposed. Such a policy would work best if coupled with an official pre-registration database, like the new Registry of Efficacy and Effectiveness Studies (REES; Spybrook, Anderson, & Maynard, 2019), a pre-registration website designed specifically for educational interventions (engagement with this registry has now become an encouraged component of new IES RFAs).

It should be noted that interventions need not affect long-run outcomes in order to be worthwhile or informative. For example, funders and researchers may find merit in a study examining the effects of a preschool reading curriculum, regardless of the curriculum's effects on long-term reading achievement. However, if such a study had only short-run goals in mind, then this should be made explicit in both the framing of the study and the stated theory of change. In this case, future long-term follow-up could be pursued only for exploratory purposes.

More often, researchers hint at predictions about the long-run importance of a particular intervention or intervention target by citing the relatively small experimental literature that has included long-run follow-up (e.g., Heckman, 2006 – cited over 3,000 times) or correlational work highlighting the predictive validity of a particular construct. Keeping with the above

example of an early reading program, the proposed intervention might frame the importance of the study by citing correlational studies showing strong relations between early reading achievement and later school success (e.g., Duncan et al., 2007- cited over 4,000 times), or they might cite influential theoretical work predicting that early boosts in reading achievement should lead to future skill acquisition (e.g., Cunha & Heckman, 2007- cited over 2,500 times). In these cases, long-term hypotheses are made implicitly, even if the study is only funded to test impacts on short-run measures of reading achievement. By asking researchers to shift these implicit theories to explicit predictions, researchers will be given incentives to think carefully about the mechanisms that connect their intervention models to the larger goals of educational programs that researchers often discuss only superficially at the beginning of papers and grant proposals.

Next, coupled with the pre-specification of long-run hypotheses, funders could also ask researchers to provide some indications for how their long-term hypotheses could be tested. By building "future research plans" into new grant proposals, funders would ask researchers to design new intervention studies that open the possibility of future, high-quality, follow-up research. Such plans could include proposed partnerships with organizations that house administrative data, or researchers could even detail plans to transfer the study to other organizations that may be better suited for future waves of data collection. As we detail in the "Recommendations for Researchers" section below, some early planning for future follow-up could substantially boost a study's chances of collecting further data from their sample should researchers and funders choose to pursue long-run follow-up.

Selecting studies for follow-up funding. If these two changes were made to the application process for new intervention studies, funders could rely on several selection mechanisms to choose from the pool of studies that 1) articulated hypotheses regarding long-run

effects, and 2) provided credible research plans for testing long-term hypotheses. First, organizations could build on the current practice of calling for follow-up of existing evaluations in RFAs. Placing new emphasis on funding follow-up studies [see recent blog post from current IES director M. Schneider: https://ies.ed.gov/director/remarks/6-19-2019.asp], even designating some RFAs entirely for follow-up funding, could encourage researchers to apply. Further, allowing researchers to apply to extend their pre-existing evaluation projects may also encourage more follow-up applications. If long-run hypotheses and research plans were already articulated in initial applications, then extension applications could be briefer and focused solely on updating the follow-up data collection plans given the current state of the research project.

Another approach could add efficiency to the process by cutting out researcher-written follow-up applications altogether if funding agencies determined themselves which projects merit follow-up. With this plan, funders would use initial grant applications to determine which studies made plausible long-term predictions and provided details for long-run data collection plans. They could then use annual progress reports to track important design issues (e.g., study attrition, implementation fidelity, etc.) to generate a pool of high quality studies eligible for further follow-up funding. Funders would then appoint a review panel to review already-funded evaluations that were nearing project completion, and they could choose which projects were most promising for follow-up based on theories of change, reported effect sizes, and design quality. Of course, with this policy, funders would merely offer funding to keep projects going, and researchers would have to consider whether accepting the funding was a worthwhile investment of their own time and energy.

Although these new funding options would offer improvements over the status quo, these mechanisms also carry drawbacks. If follow-up funding is contingent on showing "promising"

short-run effects, then researchers would be even further incentivized to design evaluations that produce the largest short-run impacts regardless of how these impacts extend into future periods. If the primary reason to follow up on an evaluation is the size of the initial impact estimate, then this positive selection (whether selection is correlated with systematic error, such as selective reporting of the largest impacts, over-aligned outcome measures, or even random error in impact estimates) will inflate end-of-treatment effect sizes. Indeed, because these incentives already exist, the current preponderance of fadeout effects in educational intervention studies (Bailey, Duncan, Odgers, & Yu, 2017) could be partly due to the fact that follow-up attempts almost exclusively ensue after "promising" short-run effects have been reported. Moreover, this preference for studies showing large short-run effects gives researchers few incentives to pursue interventions that move more difficult-to-alter aspects of student cognition and behavior. Such programs may have the best chance of producing long-lasting effects despite producing smaller short-run impacts when compared with narrowly targeted interventions.

Consequently, our preferred selection mechanism would involve funders randomly selecting projects from the aforementioned pool of high-quality studies eligible for follow-up funding. A random selection process would mitigate the incentives for researchers to design evaluations that might inflate short-term impact estimates (though the pressure to publish may still encourage many of these same behaviors), and would also incentivize more careful thinking about long-term mechanisms. The random selection process would also allow for the possibility of detecting long-term impact patterns that we have little chance of detecting in educational RCTs under the status quo (e.g., null short-term impacts followed by positive long-term impacts). Thus, randomly selecting studies that pre-specified long-term hypotheses and met a

threshold for design quality could yield substantial benefits by realigning researcher incentives and increasing the range of studies reporting long-term effects.

Prioritizing research quality. If funders encourage researchers to pre-specify plausible long-term hypotheses and future follow-up data collection plans, then the initial competitive grant review process should yield a pool of high-quality studies eligible for follow-up funding (while still funding important short-run interventions with no hypothesized long-term effects). Regardless of the specific selection mechanism pursued by funders, the field would substantially benefit if follow-up support was extended based on the quality of research, rather than the size of the short-run effect.

This could mean that funders invest in follow-up of studies that pre-specified long-run hypotheses, but found disappointing short-run effects. Funding these studies may seem risky, as analyses of long-run follow-up data would qualify as "exploratory" (i.e., any long-term effects detected would not occur due to the mechanisms pre-specified in the original theory of change). Nevertheless, many educational programs currently under consideration (e.g., public preschool, charter schools, after school programs) have been hypothesized to affect a broad range of child developmental processes, and it remains unclear whether we have fully identified, or capably measured, the mediational mechanisms that might produce long-term impacts for many of these programs. For example, in early childhood research, the famous Perry Preschool Program produced strong long-term impacts on adult indicators of economic success and well-being, yet the mediational processes that led to these impacts are still not totally understood (Bailey, Duncan, Odgers, & Yu, 2017; Heckman, Pinto, & Savelyev, 2013). Perry Preschool produced fading impacts on measures of childhood IQ, but longitudinal data collection persisted—and the study continues to yield substantial theoretical benefits as a result.

Thus, funders must determine how much emphasis should be placed on pursuing longitudinal follow-up, some of which may be exploratory. Certainly, if we find that short-run null effects are always followed by long-run null effects, then the field could learn from this and shift priorities accordingly. Even in this case, these null-effect studies would serve as an important comparison group to studies that did find positive long-term impacts. Moreover, by investing in high-quality long-run research now, we will develop an empirical body of literature that will improve our ability to rely on short-run evidence to project long-run effects in the future.

Recommendations for Researchers

If more funding is extended for follow-up studies, researchers could take advantage of these resources to enhance their intervention research in several interesting ways. First, we recommend that researchers begin planning early for potential long-term follow-up. Careful consideration of plausible long-term mechanisms from the outset of intervention development could provide substantial benefits. For example, in the above-described hypothetical reading intervention, will the curriculum teach material that students in the control group are scheduled to learn months after the end of treatment? If so, in those months, is there a plausible mechanism through which the knowledge gained during the intervention would transfer to other domains? If not, is there some way to alter the curriculum or its timing to make this more likely? Designing interventions that can purposefully connect to the set of environmental experiences expected for students after leaving the intervention would raise the possibility of developing educational interventions that will produce long-lasting effects. Researchers often attribute intervention effect fadeout to the subsequent environmental experiences of intervention participants, a possibility we find plausible. However, this possibility also points to the potential usefulness of

interventions designed to complement the subsequent environmental conditions of intervention participants.

Second, we encourage researchers to take advantage of the vast amounts of secondary data now available to continue following their evaluation samples. Penner and Dodge (2019) recently included this among the many benefits that can be gained by engaging with administrative data sources. Indeed, IES has funded multiple longitudinal data systems in states and large cities across the country (see full list at https://nces.ed.gov/programs/slds/), yet these large data systems have been largely under-utilized. Merging secondary data sources with earlier intervention evaluation samples has already yielded highly influential findings (Chetty et al., 2011; Chetty, Hendren, & Katz, 2016; Lipsey, Farran, & Durkin, 2018), and will likely continue to do so. Using administrative data sources also has the benefit of carrying a lower price than traditional modes of data collection, raising the possibility of pursuing long-term follow-up even when further funding is not guaranteed.

Given the continued growth in this sector, we encourage researchers to begin communicating with organizations that maintain administrative databases early in their intervention evaluations. This would allow researchers to better understand, and collect, the information that will be needed to eventually link participant data to secondary sources. Further, researchers should reach out to these organizations to acquire information regarding the informed consent procedures that will be required to link participant data. In some cases, it may be possible to build consent for future data release into the early waves of data collection, when participant retention and recruitment presents a less severe problem. By obtaining permission for the release of records from the outset, secondary data sources could substantially help curb long-term attrition across studies.

Of course, the benefit of these administrative sources of data should not be overstated. Such sources of data often provide measures for a narrow set of outcomes that may or may not be useful to a given study (i.e., test scores, GPA, etc.). Further, as children move out of schools or districts over time, participants may disappear from certain databases, further eroding study power. However, partnering with organizations that house higher-level databases (e.g., the state-level databases set up by IES), rather than single schools or districts, may prove valuable as participants disperse over time.

Finally, we recognize the need and desire to continue to develop new intervention projects for funders and researchers alike, and suggest that these goals can be complementary. Ongoing innovation through the development of new interventions will generate important variation that might be used to isolate effective program features. Focusing solely on older evaluation studies could have the drawback of diverting attention from the development of newer programs. One promising approach for combining these goals could be the use of older samples to test the efficacy of new programs. If both the "new" and "old" intervention were randomly assigned, testing the effects of one intervention should have no bearing on our ability to detect effects for the other. If studies were properly powered, this would also heighten our ability to find instances of "dynamic complementarity," which is the influential idea that educational investments may positively interact across time to make long-lasting impacts on children's trajectories (Cunha & Heckman, 2007). Indeed, this design has been recently pursued by at least one IES-funded project.[2]

Of course, researchers would have to consider whether older samples are representative of populations of interest for newer interventions. Further, because educational researchers often specialize in programs targeted to specific age groups (e.g., early childhood, adolescence,

transition to adulthood, etc.), providing new interventions to older samples would incentivize further collaborations between researchers across specializations. This might lead to more programs that that align instruction and programmatic elements over the course of development.

Conclusion

The field could substantially benefit from more rigorous educational evaluations reporting long-term follow-up. At present, connections between short-run outcomes and long-term impacts are often assumed, but rarely tested using experimental methods. Indeed, correlational and quasi-experimental evidence should continue to play a role in longitudinal research. However, by pursuing more longitudinal follow-up of high-quality educational RCTs, funders and researchers can better test the long-run theories that are often implied by correlational work.

Certainly, longitudinal evaluations are not without their own limitations. As longitudinal follow-up stretches into future years, the context within which the intervention was originally tested differentiates further from the status quo. This is an unfortunate, but unavoidable, limitation of longitudinal work. However, as the enduring influence of the handful of educational RCTs with long-run follow-up demonstrates (Campbell et al., 2002; Heckman, 2006; McCormick et al., 2006; Myers et al., 2004; Schochet, Burghardt, & McConnell, 2008), the underlying processes tested by interventions of interest often remain surprisingly relevant over time.

Producing long-lasting impacts on key developmental outcomes should not be considered an easy task, and the "success" or "failure" of interventions should not be judged solely on the basis of long-run effects (e.g., an intervention may be necessary, but not sufficient, for spurring long-run change on an outcome of interest). In other words, many educational programs should

probably not be expected to produce "inoculation effects." However, the common practice of citing long-run experimental or correlational evidence as motivation to pursue short-run interventions that produce unknown long-run effects indicates a need for clarity on these issues.

Thus, our longitudinal theories should be formalized and tested empirically. Perhaps researchers do expect long-run impacts of their interventions; perhaps they expect long-run impacts contingent on some measurable medium-run contextual effects; perhaps they have no specific theory in mind but merely cite long-run evidence because it is common practice to do so. Perhaps researchers refrain from discussing long-run impacts, because their educational intervention serves some worthwhile short-term goal. In any of these cases, requiring applicants to make these goals explicit would make funding decisions better informed by the purpose of the proposed research (and by reviewers' judgments of whether these goals are likely to be reached)—outcomes to which we hope funding agencies and researchers aspire.

Given the recent advancements in rigorous methodology for the evaluations of education programs, along with the new availability of administrative data sources, the opportunity for researchers and funders to support long-term follow-up has never been greater. The benefits stemming from the changes we propose would take years to accumulate, but investing in long-term follow-up projects now could yield substantial long-term benefits to the field for years to come.

Notes

1. Here, we consider traditional RCTs where the treatment group is compared with a "business-as-usual" control group. RCTs with "waitlist" control designs also disallow for long-run follow-up.
2. See recent work on the Chicago School Readiness Project (Raver et al., 2011; Watts et al., 2018), which followed an early childhood intervention sample into adolescence, and re-randomized the sample to a mindset intervention.

References

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

Bartik, T. J. (2014). *From Preschool to Prosperity: The Economic Payoff to Early Childhood Education*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. Retrieved from: https://doi.org/10.17848/9780880994835

Bartik, T. J., Gormley, W., & Adelstein, S. (2012). Earnings benefits of Tulsa's pre-K program for different income groups. *Economics of Education Review*, *31*(6), 1143-1161.

Brooks-Gunn, J. (2003). Do you believe in magic?: What we can expect from early childhood intervention programs. *Social Policy Report*, *17*(1), 1-16.

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education: Young adult outcomes from the Abecedarian Project. *Applied Developmental Science*, *6*(1), 42-57.

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

Chetty, R., Hendren, N., & Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment. *American Economic Review*, *106*(4), 855-902.

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, *97*(2), 31-47.

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*(3), 418-446.

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

Dodge, K. A., Bai, Y., Ladd, H. F., & Muschkin, C. G. (2017). Impact of North Carolina's early childhood programs and policies on educational outcomes in elementary school. *Child Development*, *88*(3), 996-1014.

Duncan, G. J., (2015, March). *Fade-out in human capital intervention: Death, miracles and resurrection.* Lecture conducted for annual meeting of the Society for Research on Educational Effectiveness, Washington, D.C.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., … & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428-1446. doi: http://dx.doi.org/10.1037/0012-1649.43.6.1428.supp.

Fredriksson, P., Öckert, B., & Oosterbeek, H. (2013). Long-term effects of class size. *The Quarterly Journal of Economics*, *128*(1), 249-285.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*(5782), 1900-1902.

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.

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. *Yearbook of the National Society for the Study of Education*, *104*(2), 99-118.

Kraft, M. A. (2018). *Interpreting Effect Sizes of Education Interventions*. Brown University Working Papers). Providence. Retrieved from https://scholar.harvard.edu/files/mkraft/files/kraft_2018_interpreting_effect_sizes.pdf

Krueger, A. B. (2003). Economic considerations and class size. *The Economic Journal*, *113*(485), F34-F63.

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.

Martin, J., McBridge, T., Brims, L., Doubell, L., Pote, I., & Clarke, A. (2018). *Evaluating early intervention programmes: Six common pitfalls, and how to avoid them.* Retrieved from Early Intervention Foundation website: http://www.eif.org.uk/publication/evaluating-early-intervention-programmes-six-common-pitfalls-and-how-to-avoid-them

McCormick, M., Hsueh, J., Weiland, C., & Bangser, M. (2017). *The challenge of sustaining preschool impacts.* Retrieved from MDRC website: https://www.mdrc.org/publication/challenge-sustaining-preschool-impacts

McCormick, M. C., Brooks-Gunn, J., Buka, S. L., Goldman, J., Yu, J., Salganik, M., ... & Bauer, C. R. (2006). Early intervention in low birth weight premature infants: results at 18 years of age for the Infant Health and Development Program. *Pediatrics*, *117*(3), 771-780.

Myers, D., Olsen, R., Seftor, N., Young, J., & Tuttle, C. (2004). *The impacts of regular Upward Bound: Results from the third follow-up data collection*. Washington, DC: Mathematica Policy Research.

Penner, A. M., & Dodge, K.A. (2019). Using Administrative Data for Social Science and Policy. *RSF: The Russell Sage Foundation Journal of the Social Sciences, 5(2),* 1–18. doi: 10.7758/RSF.2019.5.2.01.

Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., … Weiland, C. (2017). *The Current State of Scientific Knowledge on Pre-Kindergarten Effects* Retrieved from Brookings website: https://www.brookings.edu/wp-content/uploads/2017/04/duke_prekstudy_final_4-4-17_hires.pdf

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

Schneider, M. (2019, June 19). Some Thoughts on the New IES RFAs [Blog post]. Retrieved from https://ies.ed.gov/director/remarks/6-19-2019.asp

Schochet, P. Z., Burghardt, J., & McConnell, S. (2008). Does job corps work? Impact findings from the national job corps study. *American Economic Review*, *98*(5), 1864-86.

Singer, J., (2019, March). *Shaping the arc of educational research.* Hedges Lecture conducted for annual meeting of the Society for Research on Educational Effectiveness, Washington, D.C.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 5-14.

Spybrook, J., Anderson, D., & Maynard, R. (2019). The Registry of Efficacy and Effectiveness Studies (REES): A Step Toward Increased Transparency in Education. *Journal of Research on Educational Effectiveness*, 12(1), 5-9.

Stipek, D., Franke, M., Clements, D., Farran, D., & Coburn, C. (2017). PK-3: What Does It Mean for Instruction? Social Policy Report. Volume 30, Number 2. *Society for Research in Child Development*. Retrieved from www.srcd.org/publications/social-policy-report

Watts, T. W., Gandhi, J., Ibrahim, D. A., Masucci, M. D., & Raver, C. C. (2018). The Chicago School Readiness Project: Examining the long-term impacts of an early childhood intervention. *PLOS ONE*. doi: 10.1371/journal.pone.0200144

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, *84*(6), 2112-2130.