Monograph Matters

SRCD Society for Research in Child Development

Nguyen, T. & Bailey, D. H. (2021). *Drawing General Conclusions from Null Effects of a Prekindergarten Curriculum: Challenges and Solutions*. [Peer commentary on the article "Effects of prekindergarten curricula: *Tools of the Mind* as a case study" by K. T. Nesbitt and D. C. Farran]. *Monograph Matters*. Retrieved from https://monographmatters.srcd.org/2021/02/16/commentary-nguyenbailey-86-1/

Drawing General Conclusions from Null Effects of a Prekindergarten Curriculum: Challenges and Solutions

Tutrang Nguyen
Mathematica
tnguyen@mathematica-mpr.com

Drew H. Bailey
University of California, Irvine

In their monograph, Nesbitt and Farran (2021) use the *Tools of the Mind* (hereafter, *Tools*) curriculum as a case study for evaluating the effects of intentional, scripted prekindergarten curricula. Despite expectations of the developers of *Tools* that the prekindergarten curriculum would have positive effects on children's academic, executive function (EF), self-regulation, and social skills through the end of first grade, Nesbitt and Farran find no positive effects of *Tools* on any of these outcomes. These null effects are policy relevant.

However, the generalizability of these findings to higher-order constructs (e.g., other implementations of *Tools*, other early childhood curricula) and other contexts (e.g., kindergarten classrooms, geographical regions) is difficult to judge given the many potential theory- and program-related explanations for the findings. Nesbitt and Farran built in an unusually strong set of design and analytic features that improve the field's understanding of the plausible explanations for these results. In this commentary, we describe these explanations and assess where challenges might lie when attempting to draw conclusions from null findings in field experiments.

Technical Challenges

Learning from null impacts in field experiments can be difficult. First, field experiments in education tend to be far messier than lab experiments, thus making null effects more difficult to interpret. Consider a hypothetical pair of experiments, both comparing the effectiveness of a new kind of memory practice, but one implemented in a lab setting and the other implemented in a school classroom.

In the lab experiment, participants are randomly assigned to receive or not receive a novel type of memory practice, and the experimenter is likely to exert strong control over the amount of time participants are exposed to the practice strategy. In such experiments, it is simple to attain nearly perfect agreement among assignment to treatment, delivery of the treatment, and student adherence to the treatment. Null impact of the treatment can then be interpreted as theoretically meaningful. For example, a lab study like this showing no better memory in

children assigned to the treatment condition would allow one to conclude that the tested treatment (at least at the studied dosage) has no impact on memory performance.

In contrast, in a field experiment, a researcher might provide training to some teachers about how to implement the same form of memory practice. Assuming a teacher is assigned to the treatment group, the student may or may not actually receive the instruction intended by the program designers. Or if they receive the instruction, several students in the control group may receive similar kinds of instruction in their classrooms during the implementation period, weakening the treatment-control contrast on classroom practice. Among teachers assigned to the treatment condition, some may not adhere to the assigned methods if they do not prefer those curriculum methods to their previous methods, or if they lack adequate supports to learn and adhere to the curriculum. Therefore, results can be null assuming large influences of assignment on receipt (in this case, of the professional development (PD)), receipt on adherence (to the desired type of instruction), or adherence on learning, assuming at least one of these effects is small. For these reasons, and in contrast to our hypothetical lab experiment, it can be difficult to tell whether a null result of this field experiment indicates a failure of theory, assignment, receipt, or adherence.

Nesbitt and Farran's evaluation does an unusually strong job of addressing these challenges to understanding null findings in field experiments. One interesting and unique aspect of their monograph is the clear description of the involvement of the curriculum developers in the evaluation study. This involvement is important because having the developers involved in the research preempts certain critiques that frequently arise in debates about evaluation studies. That is, if effects are weak or absent, the developers may suggest that training duration was inappropriate, too sparse, or that the wrong outcomes were measured. The inclusion of curriculum developers thus makes for a stronger test of the program. We commend the authors for including various methods to improve implementation in their study, such as working directly with the *Tools* developers and trainers to get them to develop hypotheses about what classroom processes they thought would differ in treatment classrooms compared with business-as-usual classrooms.

Nesbitt and Farran examined whether problems of implementation were a plausible explanation for null impacts. In Chapter III, they examined the extent to which treatment teachers implemented specific instructional practices prescribed by *Tools* and whether the implementation of these practices was related to children's gains in academic, EF, and social skills. They found no consistent associations between (a) the amount of time the curriculum was implemented and the fidelity with which teachers followed through with the activities and (b) children's gains in any of these domains.

Together, these results might suggest a possibility of a theory or program (rather than implementation) failure: the authors acknowledge that their evaluation raises "more general questions about how curriculum experiences manifest themselves in assessed skills" (p. 50). The kinds of activities included in *Tools*—at least at typical levels of dosage—may not be well-designed to strongly influence preschoolers' academic, EF, self-regulation, and social skills. The viability of this potential interpretation is strengthened by Nesbitt and Farran's literature review of studies summarized in Table 1. They conclude that the six previous randomized control trials (RCTs) of *Tools* implemented in preschools "found limited evidence that the *Tools*

prekindergarten curriculum has significant positive impacts on children's academic, self-regulation, or socio-emotional skills" (p. 28).

The Challenge of Changing Minds

We find Nesbitt and Farran's conclusions to be mostly convincing. However, the study also led us to consider some difficult and more general questions about what can be learned from null impacts in program-evaluation studies.

First, after six RCTs of *Tools* in preschools with limited evidence of impacts on the targeted child outcomes, to whom *should* it be a surprise that the current evaluation generated largely null impacts? The authors note that curriculum developers with whom they were working expected to see impacts of *Tools* on outcomes being tested in the current research. This expectation implies that the developers attributed previous null results to incomplete or faulty program implementation.

Second, what will *Tools* optimists think after this new evaluation? When the designers and trainers of the intervention and evaluators are different people, these parties might be prone toward different explanations of null findings. Designers, prone to thinking that their interventions work, may be averse to explanations that point to small effects of adherence to the intervention on key outcomes. Evaluators, prone to thinking that the evaluation was informative, may be averse to explanations that point to implementation failures.

For the moment, we will attempt to take the perspective of a reasonable advocate for the *Tools* program. (To be clear, we are not advocating this position, but are instead attempting to anticipate how an advocate might respond to the data and arguments reported by Nesbitt and Farran). From this assumed position, we might argue that fidelity was qualitatively different from what we had expected, as evidenced by the null impacts on child talk and smaller-than-predicted impacts on other classroom process outcomes. Further, we might claim that the null relations between classroom-process outcomes and child learning might primarily reflect measurement error in the former. Of course, in an imagined conversation between the advocate for *Tools* and the evaluators, the evaluators could object, noting that the designers agreed that these process-measures would likely be affected by the PD as it was delivered. They might also note that small associations between these practices and child outcomes suggest increases in implementation fidelity would not have led to substantially larger impacts. The *Tools*-advocate might argue that something unmeasured about the implementation is responsible for the lack of correspondence between PD and impacts, and the evaluators could respond that this argument is unconvincing, because it was made by the *Tools*-advocate in hindsight, after seeing the dearth of impacts on key outcomes.

Such arguments raise the pessimistic prospect that null impacts from field studies may rarely sway the opinions of important actors in educational practice and policy, even when studies are carefully designed to learn from them. With this perspective in mind, and inspired by Nesbitt and Farran's care to improve the understanding of null impacts from a field RCT, we propose an inexpensive method for collecting additional process data specifically designed to increase the possibility that evaluations showing null impacts will have the potential to change the minds of educational researchers and practitioners who are optimistic about programs at the start of the evaluation.

Addressing Social and Psychological Challenges of Learning from Null Impacts

We offer a suggestion for future data collection that might reduce the possibility of hindsight bias, improve the potential for such work to change the minds of educational researchers and practitioners, and provide useful additional process information that might be missing from existing observation protocols. In brief, we suggest that future field experiments that include both designers and evaluators of an intervention in collecting dynamic forecasts from both parties before and throughout the evaluation. Forecasts have been advocated for improving the quality of policy debates and for adding useful information about the likely effectiveness of interventions (DellaVigna, Pope, & Vivalt, 2019; Tetlock, Mellers, & Scoblic, 2017). Preregistered forecasts are a useful tool for protecting against hindsight bias by making it clear to all parties what findings were unexpected. Quantitative forecasts can be useful for forecasting the impacts of experiments (Dreber et al., 2015), including interventions (DellaVigna & Pope, 2018). They also allow for the possibility for designers or evaluators to respond to unexpected factors identified during implementation, but before having already viewed experimental impacts.

We propose that the two teams—one of curriculum developers and the other of evaluators—would be asked to make quantitative forecasts about the impacts of the treatment on key intervention outcomes. Forecasts should begin early, starting when proposals are first submitted for funding, where they could be used for power analyses. At each major point during the implementation of the evaluation, teams should be asked to provide updated forecasts. The specific stages might differ depending on the design of the evaluation, but these might commonly be scheduled (1) after the first set of observations during a PD intervention; (2) after the initial observational data are analyzed and presented to the designers and evaluators; (3) after the first set of classroom observations during implementation have been collected, analyzed, and presented; and (4) after the intervention is complete but before impacts have been estimated and disseminated to the broader community of researchers, practitioners, and policy makers.

Possible additional features that might strengthen the usefulness of these forecasts include allowing either team to update forecasts based on any other new information they have received (e.g., reports from participants; additional observations; new research released during the program period) and accompanied by qualitative justifications for these updates. Researchers otherwise unaffiliated with the evaluation project could be asked for forecasts as a test on whether the theory of change is promising at the start of the evaluation or not. To discourage gaming forecasts (e.g., by strategically over-estimating at the start of the study and dropping forecasts to a more realistic level during implementation) and to discourage over-adjusting one's forecasts on the basis of any single event, researchers' forecast accuracies should be scored as deviations between realized impacts and forecasts throughout the study (i.e., not just at the final forecasting period). A team that makes reliably poor forecasts across studies might suffer some reputational damage.

Dynamic forecasts allow for additional tests of where and when failures take place. If the design or research teams drop their forecasts during PD, this would suggest a different problem than if the largest drop occurs during implementation. If drops occur after seeing process data (and these dropped forecasts are later found to be better matched to later observed impacts), this might provide useful validity evidence for later measured forecasts. If the design team keeps

forecasts high until seeing null impacts, the attribution to implementation-failure is less convincing.

Throughout the monograph, Nesbitt and Farran describe their close collaboration with the *Tools* developers, both during the proposal stage and during the actual project activities: "All assessments, observation measures, and fidelity systems were chosen in collaboration with the developers and all were beta-tested with national *Tools* trainers before final decisions were made" (p. 97). In Chapter IV, the authors examine the *Tools* developers' own hypotheses about whether the curriculum affected general classroom processes. But it was unclear to us what Nesbitt and Farran themselves had hypothesized. We found that there was no mention of instances in which the evaluators might have disagreed with the developers during various stages of the study. Trying to learn about such disagreements after the fact would be difficult, both because of the curse of hindsight and because disagreements are not always fully communicated during a collaboration. Systematically and prospectively documenting how the evaluators and developers generated their hypotheses, how they disagreed and negotiated study decisions, or even when they were not necessarily on the same page would have provided useful additional information. This information would benefit education research directly by providing important information that can be used to help understand went wrong and when. To be clear, Nesbitt and Farran already involved the developers substantially in the process; it is only because they were so involved that we have considered various ways in which the designers and evaluators might have changed their views about the evaluation at various points during the study.

Incorporating dynamic forecasts into program evaluations has several potential indirect benefits as well. First, committing to a forecast makes surprise more likely, and surprise can improve the forecaster's learning. If a developer forecasts an impact of 0.40 SD on an achievement test right up until the impacts are estimated from the data on a key outcome, but the observed impact is approximately zero, it is difficult to claim, after the fact, that the null program impact was obvious based on inadequate implementation. That is, if the designers forecast positive impacts based on fidelity data, along with their own judgment based on observations of the training, classrooms, and other available information, then implementation concerns expressed after seeing impacts are less convincing than had the designers been able to forecast null impacts during implementation before data have been analyzed. Symmetrically, if an evaluator's forecast is substantially higher than the developer's forecast until the null impacts are estimated, this would give more weight to the possibility that difficult-to-measure aspects of implementation can be perceived by developers and hopefully convince the evaluator that the evaluation was not a strong test of the program's efficacy.

Another indirect benefit of integrating dynamic forecasts into a program evaluation is that researchers' forecasts can be used by the researchers and others to improve their and others' future work. If researchers analyze and come to understand their forecasting errors and successes, their future forecasts may become better calibrated to the likely impacts of programs across outcomes, leading them to invest their time and effort more strategically into more promising programs. If a research or development team shows strong forecasting skills, they might gain influence for their demonstrably strong insights.

Better understanding what factors lead to better versus worse forecasts will be useful for the field of program evaluation. What kinds of biases do evaluators and designers typically show?

How might we incorporate this information into the design of better programs (e.g., by understanding at what point(s) in the theory of change forecasts drop toward an eventual null impact; do forecasts drop more during the evaluation of an efficacy or effectiveness trial; what kinds of programs do researchers forecast will show the strongest impacts on broad outcomes?) and evaluations (e.g., to improve receipt or adherence)? If researcher forecasts at various evaluation stages are found to be reliable, perhaps when forecasts drop under a predetermined threshold, they could be used as a warning sign to consider ending or modifying evaluations early to save resources. Of course, such a policy would involve several tradeoffs: we are not sure whether it would be better to make forecasts hidden to preserve independence of forecasts and prevent conflict between the teams or to make them open to facilitate better implementation on the fly. The efficiency and ethics of such decisions will rely on emerging information about the accuracy of forecasts. For example, if forecasts are unreliable, they may be more useful for effectively surprising and changing researchers' minds than for informing study design in real time. Decisions should also depend on the goals of the particular evaluation. For example, if a program is already being implemented at scale, and part of the purpose of the evaluation is to measure levels and variation in implementation without altering it during the evaluation, then perhaps designers and evaluators should make forecasts as independently as possible; in contrast, if a program is being designed and iteratively improved, forecasts could be shared throughout the evaluation as part of the improvement process.

Conclusion

Nesbitt and Farran present the results of an evaluation of the *Tools* of the Mind curriculum, convincingly arguing that as a preschool curriculum, the program is unlikely to reliably generate substantial impacts on children's school readiness. We applaud their careful design and analysis, and we highlighted some of the useful components of their work. They moved beyond simply reporting the impacts of *Tools* on children's outcomes and probed further for what might have led to the null results. The authors focused on teachers who had been randomly assigned to the treatment condition, and examined the degree to which the curriculum was delivered as intended and the connections between fidelity of implementation and children's outcomes in prekindergarten. They also examined curriculum effects on classroom processes and investigated whether various classroom processes were associated with gains in children's outcomes. We hope their thorough evaluation will change the thinking of educational researchers and practitioners.

At the same time, however, we worry that program-evaluation studies do not change minds as effectively as they might. In our commentary, we proposed an inexpensive method for incorporating dynamic forecasts into a program evaluation as an additional indirect measure of process. We suggest that a widespread and longstanding implementation of this proposal might ultimately extend our ability to distinguish between theory or program vs. implementation failures, to change the minds of more readers when the evidence warrants. Further, we suggest that a body of dynamic forecasting research would help designers, evaluators, and funders to think about how to better select promising projects, along with where particular types of programs are most likely to fail.

References

DellaVigna, S., & Pope, D. (2018). Predicting experimental results: Who knows what? *Journal of Political Economy*, *126*(6), 2410-2456. https://doi.org/10.1086/699976

DellaVigna, S., Pope, D., & Vivalt, E. (2019). Predict science to improve science. *Science*, *366*(6464), 428-429. https://doi.org/10.1126/science.aaz1704

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., ... & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343-15347. https://doi.org/10.1073/pnas.1516179112

Nesbitt, K. T., & Farran, D. C. (2021). Effects of prekindergarten curricula: *Tools of the Mind* as a case study. *Monographs of the Society for Research in Child Development*, 86(1). https://doi.org/10.1111/mono.12425

Tetlock, P. E., Mellers, B. A., & Scoblic, J. P. (2017). Bringing probability judgments into policy debates via forecasting tournaments. *Science*, *355*(6324), 481-483. https://doi.org/10.1126/science.aal3147