BROOKINGS

SERIES: The Brown Center Chalkboard | Number 56 of 93

Paper | February 26, 2014

# Does Pre-k Work? It Depends How Picky You Are

By: Grover J. "Russ" Whitehurst

**When I headed the Institute of Education Sciences in the U.S. Department of Education, I was sometimes called into meetings organized by the Office of the Secretary that were responsive to requests by organizations lobbying the Department's front office. In the early days of No Child Left Behind, these organizations felt compelled to justify their pitch by citing research (what with the phrase "scientifically-based research" occurring 111 times in the NCLB statute). We would sit through the dog and pony show, and then one of the Secretary's people would turn to me to be the bad cop by explaining why the research that had been invoked in the presentation wasn't credible. I remember a meeting in which one of my colleagues, trying to soften my blow more than I had been able to, told our guests in her sweet-as-honey Texas drawl, "What Russ is trying to say is that ya'll just need to be more picky."**

My recent writings and presentations on early childcare have been motivated by what I see as the weak evidence behind the groundswell of advocacy for public investments in statewide universal pre-k. Opponents of my position have appealed, just as I have, to findings from scientific research. Assuming good faith and honest communication, how is it that different individuals could look at the same research and come to such different conclusions?

The standards I apply are in keeping with those established by the What Works Clearinghouse (through which the U.S. Department of Education vets research on the effectiveness of education programs and products). The WWC standards,[i] in turn, are broadly consistent with those used by many federal agencies, including the FDA, to judge evidence of effectiveness. And they are standards that align with the top tier of the Obama administration's tiered-evidence approach.[ii]

What are these standards? First, they concern the internal validity of the evaluation of the program in question. To have high internal validity the evaluation must demonstrate beyond reasonable doubt that the program in question had a causal impact on the outcomes it was intended to influence. This is best accomplished by findings from one or more well-designed and implemented randomized controlled trials (RCTs). RCTs are the gold-standard for evaluating the effectiveness of social programs because the act of randomly assigning participants to the program or control group assures that, to a statistically determinable margin of error, the two groups are identical on everything that could influence the outcomes being

measured except their group assignment. Methods other than RCTs can provide useful information on what works, but they always necessitate more assumptions than RCTs. And while these other methods sometimes produce results that are very close to those produced by RCTs, they also often produce results that are demonstrably wrong when compared with the results of an RCT. That is why the RCT is the scientific bedrock for determining whether social programs work. Anything else is what you do when an RCT isn't possible.

The second component of high quality evidence standards is external validity (or relevance). Findings from evaluations with high external validity have been obtained in settings that are representative of those in which the program is to be implemented, and address whether differences between the outcomes for program and control groups are sufficiently sizable and sustained to make them worthwhile with respect to the program's goals.

With that background, I'll take you quickly through my ratings of the internal and external validity of the studies that have been most frequently cited in the debate on the Obama administration's proposal for Preschool for All and related efforts. There are dozens of other studies that I might have included, but I don't have the space to do so and their inclusion wouldn't lead to different conclusions.

## Programs from the 1960s and 1970s

| Program/Research | Reported Impact (after initial year) | Internal Validity | External Validity |
|---|---|---|---|
| Perry Preschool | + | A- | C |
| Abecedarian | + | B+ | C |
| Chicago Child Parent | + | C | B |
| Head Start in the 1960s | + (for mortality) | B | C |

Evaluations of the four programs listed in the table above all report long-term positive impacts for participants in early childhood education programs, e.g., higher test scores in school or less criminal activity in adulthood. But they are wanting in terms of their external validity for decisions about whether to expand present public programs for four-year-olds: They are from a time when very little of today's safety net for the poor was in place, when center-based care for four-year-olds was rare and even kindergarten was not the rule, and before the wave of Hispanic immigration that transformed the demographics of early education programs for children from low-income families. Further, all but Head Start in the 1960s were multiyear intensive interventions rather than one year programs, and two of the four (Perry and Abecedarian) were small single-site programs run by their developers. Concluding that findings from these studies demonstrate

that current and contemplated state pre-k programs will have similar effects is akin to believing that an expansion of the number of U.S. post offices today will spur economic development because there is some evidence that constructing post offices 50 years ago had that effect.

There are serious issues with the internal validity of these studies as well. None of the evaluations of these older programs was a well-implemented randomized trial. The highest score on internal validity goes to Perry because of the recent effort by James Heckman to repair statistically the assignment errors made by the original research team.[iii] The Perry researchers violated the rules of random assignment for an RCT in multiple ways, including shifting families that had been assigned randomly to the program group to the control group because the mothers worked and thus couldn't participate in the home component of the program. Abecedarian[iv] also had compromised random assignment. The research team assigned families to the program and control conditions before informing the children's mothers of the time commitment that would be required for mothers in the program group. At that point, a significant proportion of mothers assigned to the Abecedarian program dropped out compared to only a small proportion of mothers assigned to the control group (in which the requirements of participation were minimal). Heckman was able to carry out a statistical fix of the mistakes in random assignment made in Perry because all the children were pretested and followed. However, the early dropouts in Abecedarian were not pretested and followed, so no statistical adjustment is possible. This leaves significant questions about whether the program families in Abecedarian were different from the control families at the outset of the study.

The Chicago Child Parent program[v] is a center-based intervention conducted at scale in Chicago, which gives it an advantage in external validity compared to small single-site programs such as Perry. But the control group was not formed through random assignment. Nor were the children in the program and control groups pre-tested and shown to be equivalent prior to program onset. Thus, the internal validity of the evaluation is weak because one cannot rule out the possibility that later differences between the two groups were simply a reflection of differences in families and children that existed prior to program. Strong external validity is undermined by weak internal validity, so not a lot of stock can be placed in findings from the Chicago Child Parent program.

The study of Head Start in the 1960s[vi] is a retrospective analysis of outcomes such as mortality and high school graduation rates, reported at the county level, comparing the very poorest counties in America, which received federal grant writing assistance for the initial round of Head Start funding, with slightly less poor counties, which did not receive such assistance. This study lacks any information on individual children and their attendance in Head Start, how funds were spent, or anything else that would ordinarily be considered the treatment of interest. Thus, the finding that mortality rates were subsequently lower in the counties receiving Head Start grant writing assistance requires a long series of linked assumptions to justify the conclusion that this had anything to do with Head Start. And the impacts found, most prominently on health, were not found in the recent National Head Start Impact Study, which has much higher internal and external validity. The lack of congruence in the findings for health should not be surprising because the health

supports for poor families that are available today in forms such as Medicaid, WIC, and food stamps were not present 50 years ago. Thus, the external validity for present day programs of the finding that an intervention from decades ago may have impacted health is questionable.

## Programs from the 1980s

| Program/Research | Reported Impact (after initial year) | Internal Validity | External Validity |
|---|---|---|---|
| Head Start in the 1980s | + | C | A |
| Infant Health and Development | + (impacts only for disadvantaged children with close to normal birth weights) | A | B |

The study of Head Start in the 1980s[vii] compares siblings within the same family — one child of which went to Head Start and one or more others either stayed home or attended another type of preschool. The study uses a large, nationally-representative dataset and thus gets an A for external validity. Siblings who were reported by their parents to have attended Head Start did better later in life than their siblings who did not attend Head Start. However, to accept that the differences in outcomes of the two groups are due to Head Start requires the assumption that the Head Start attendees and their sibling controls were equivalent except for Head Start attendance. But it seems overwhelmingly likely that a parental decision to send one child to Head Start and keep another child at home was made precisely because there were differences in the children that the parents recognized, e.g., one seemed ready for pre-k and the other not. Outcomes later in life favoring the siblings who went to Head Start could just as easily be caused by preexisting differences in children as to Head Start participation, thus the C grade for internal validity.

The Infant Health and Development program[viii] involved an intensive intervention from birth to age three for low-birth-weight children. It was a well-implemented randomized trial, thus the A grade for internal validity. It gets a grade of B for external validity because of the difficulty of generalizing results from an intensive birth-to-three intervention for low-weight infants to universal pre-k for four-year-olds. Notice from the table that positive long term results were only obtained for children from disadvantaged families who were at the high end of the low birth weight dimension. There were no impacts for children from non-poor families regardless of birth weight. This study provides the strongest evidence available on the greater return on investment of targeted preschool interventions in contrast to universal programs in which money is spent on all children, with the limitations on external validity I've described.

## Recent programs

| Program/Research | Reported Impact (after initial year) | Internal Validity | External Validity |
|---|---|---|---|
| Head Start | None | A | A |
| District Programs, e.g., Tulsa | Unknown (research design doesn't allow follow-up after pre-k) | B | B |
| Georgia & OK Universal | + ( very small at best) | B | A |
| Tennessee Voluntary Pre-K | - | A- | A |

Finally, let's consider some recent programs. The National Head Start Impact Study[ix] is one of the strongest evaluations of a social program in the last 50 years. There were a small number of outcomes favoring Head Start participants at the end of the Head Start year, but no appreciable differences between children attending and not attending Head Start from kindergarten through third grade. The study is a randomized trial, is nationally representative of Head Start centers, and includes follow-up of the sample through the end of third grade. Further, it is an evaluation of a scaled-up program for four-year-olds that is similar in most respects to the statewide universal pre-k programs that are being touted by pre-k advocates today, so it has high external validity.

There are three studies[x] of district level pre-K programs that have received considerable attention, one of Tulsa, another of the Abbott Districts in New Jersey, and another in Boston. I've written previously about methodological flaws in these studies, the most important being that researchers compared children who successfully completed the pre-K program and were just entering kindergarten (the program group), with children who were just starting the pre-k program (the control group), adjusting statistically for the age difference in the two groups. The problem with this design is that all the children who did not make it successfully through pre-k because they dropped out or moved are absent from the program group, which is tested at entry into kindergarten, whereas all the children who will eventually experience conditions that lead them to drop out are still in the control group. This means the two groups are imbalanced at the outset on factors that could well influence the outcomes. These studies also are weak with regards to external validity because the research design does not permit a determination of whether the pre-k program improves performance in elementary school — the control group begins to receive the pre-k program just after the children are initially tested, which means there is no untreated control group with which to benchmark performance in kindergarten and thereafter.

The studies[xi] of the Georgia and Oklahoma universal pre-k programs get the highest grade for external validity because the programs in these two states have been held up by President Obama and others as models to be replicated across the nation. The evaluations are based on comparison of gains in NAEP

scores in Georgia and Oklahoma before and after the introduction of universal pre-k vs. gains during the same time periods in states that did not introduce universal pre-k. There are many challenges to attributing differences that emerge across the states in NAEP gains to the presence of universal pre-k rather than to the many other ways that states differ in their policies and circumstances. Thus, the studies get a grade of B for internal validity. This means that the studies may be over- or underestimating the impact of universal pre-k on later academic performance. Advocates of universal pre-k who wish to ground their position in research better hope the estimates are biased downward because they are very small, e.g., no more than one to three percent of a standard deviation difference between the children in Georgia/Oklahoma vs. other states on fourth grade NAEP achievement scores. This is less than a one point difference on a NAEP scale on which the achievement gap between whites and blacks or whites and Hispanics is 25-30 points.

Finally, the recent evaluation of the Tennessee Voluntary Pre-K Program[xii] gets an A- on internal validity. It was designed as a randomized trial and as such should get an A, but the results reported by the research team for achievement outcomes in kindergarten and first grade exclude children who either won the lottery to attend the state pre-k program but did not attend or managed to get themselves into the state pre-k program even though that were not a lottery winner. Such a “treatment-on-treated” analysis typically produces larger effect estimates than an analysis that strictly honors the initial random assignment of participants to conditions (“intent-to-treat”). But that is not a necessary outcome depending on how the treatment-on-treated analysis is conducted. Thus we can’t be sure that the findings as reported are the same as those that would have been obtained from an intent-to-treat analysis. The Tennessee study gets an A on external validity because it is an evaluation of a current statewide pre-k program that has most of the attributes that are listed by pre-k advocates as the critical features of high quality programs. The evaluation findings are very similar to those from the Head Start Impact Study, i.e., outcomes favoring the program group at the end of the pre-k year, but no differences later in elementary school. In the case of the Tennessee evaluation, results at the end of first grade tend to favor those in the control group.

## What does the research say?

The previous tables and descriptions refer to 13 separate studies (including 3 similar studies of district programs and two similar studies of statewide programs in Oklahoma and Georgia). Of these 13, six report enduring and meaningful impacts beyond the pre-k year, four report null, negative, or very small positive impacts beyond the pre-k year, and three do not report findings beyond the pre-k year.

It would be easy for someone without the training to carefully evaluate these studies or someone with a strong motive to advocate for the expansion of publicly funded pre-k to summarize this research by saying that the *preponderance of evidence* supports universal pre-k for four-year-olds. After all, of the 10 studies I’ve reviewed that have long-term follow-up, 60 percent report substantive positive outcomes.[xiii] Libby Doggett, the Obama administration’s point person on Preschool for All, has been singing exactly this song at every opportunity:

> *You have to look at the preponderance of the evidence. Better high school graduation rates, social and emotional stability, less crime and other results speak for themselves.*[xiv]

But results do not speak for themselves. Rather, it is the combination of results and the research designs that produce them that do the speaking. And some of the combinations speak a lot louder than others.

Not one of the studies that has suggested long-term positive impacts of center-based early childhood programs has been based on a well-implemented and appropriately analyzed randomized trial, and nearly all have serious limitations in external validity. In contrast, the only two studies in the list with both high internal and external validity (Head Start Impact and Tennessee) find null or negative impacts, and all of the studies that point to very small, null, or negative effects have high external validity. In general, a finding of meaningful long-term outcomes of an early childhood intervention is more likely when the program is old, or small, or a multi-year intervention, and evaluated with something other than a well-implemented RCT. In contrast, as the program being evaluated becomes closer to universal pre-k for four-year-olds and the evaluation design is an RCT, the outcomes beyond the pre-k year diminish to nothing.

I conclude that the best available evidence raises serious doubts that a large public investment in the expansion of pre-k for four-year-olds will have the long-term effects that advocates tout.

This doesn't mean that we ought not to spend public money to help families with limited financial resources access good childcare for their young children. After all, we spend tax dollars on national parks, symphony orchestras, and Amtrak because they make the lives of those who use them better today. Why not childcare?

It does mean that we need public debate that recognizes the mixed nature of the research findings rather than a rush to judgment based on one-sided and misleading appeals to the preponderance of the evidence.

To Representative George Miller and others who think that raising questions about the quality of the research on the long-term effects of pre-k is tantamount to being a "childcare denier,"[xv] I say:

Ya'll just need to be more picky.

---

[i] http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19

[ii] http://www.whitehouse.gov/blog/2013/12/17/building-evidence-base-what-works

[iii] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3524308/

[iv] http://static.vtc.vt.edu/media/documents/245_-_Adult_outcomes_as_a_function.pdf

[v] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3793348/

[vi] http://www.crocus.georgetown.edu/reports/CROCUSworkingpaper7.pdf

[vii] http://www.people.fas.harvard.edu/~deming/papers/Deming_HeadStart.pdf

[viii] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2196752

[ix] http://www.acf.hhs.gov/sites/default/files/opre/head_start_report.pdf

[x] http://www.crocus.georgetown.edu/reports/CROCUSworkingpaper1.pdf

http://www.childrensfutures.org/Docs/AbbottPreschoolProgramStudy.pdf

http://onlinelibrary.wiley.com/doi/10.1111/cdev.12099/abstract

[xi] http://www-siepr.stanford.edu/Papers/pdf/08-05.pdf
http://www.brookings.edu/~/media/Projects/BPEA/Fall%202013/2013b%20cascio%20preschool%20education.pdf

[xii] https://my.vanderbilt.edu/tnprekevaluation/files/2013/10/August2013_PRI_Kand1stFollowup_TN-VPK_RCT_ProjectResults_FullReport1.pdf

[xiii] There are several other scaled-up preschool programs that have been evaluated with well-implemented RCTs and found to have no immediate effects, including Even Start, the Comprehensive Child Development Program, and over a dozen curriculum improvement efforts pitting model curricula against run-of-the-mill programs (see PCER and CLIO). Inclusion of these evaluations would tilt the preponderance of evidence in a different direction but would leave unchanged the general conclusion that the evidence is mixed.

[xiv] http://www.politico.com/morningeducation/1213/morningeducation12380.html

[xv] http://www.youtube.com/watch?v=fAK796qPOa4

**AUTHOR**

Grover J. "Russ" Whitehurst
Director, Brown Center on Education Policy
Senior Fellow, Governance Studies


*Read more on the challenges facing the American education system and practical policy solutions from the Brown Center on Education Policy »*