Chapter 4

Criticisms of the Research Evidence

The research evidence for early childhood education has been questioned by critics, including the Wall Street Journal editorial page, Russ Whitehurst of the Brookings Institution, Charles Murray of the American Enterprise Institute, and Shikha Dalmia and Lisa Snell of the Reason Foundation. These criticisms have been repeated in news coverage by FactCheck.org and National Public Radio.

These critics raise the following issues:

• The evidence for early childhood education is argued to be based on small experiments conducted a long time ago by researchers; therefore, critics contend, we don’t know whether early childhood education will work if run today at a large scale by ordinary public agencies.

• The recent Head Start experiment is argued to show that test score benefits of pre-K fade to statistical insignificance by third grade.

• Data from Oklahoma and Georgia are used to argue that universal pre-K fails to significantly improve test scores or other social indicators.

• A random assignment study of Tennessee’s pre-K program is used to argue against the effectiveness of state and local pre-K programs.

These criticisms are overblown. As argued below, none of these criticisms successfully challenge the research consensus: High-quality early childhood education works.

CRITICISM 1: The evidence for early childhood education’s benefits only comes from small and expensive experiments run a long time ago.
Critics acknowledge that evidence for early childhood education is provided by the Perry and Abecedarian experiments. But this evidence is argued to be irrelevant to policy today:\textsuperscript{32}

\textit{Costs per participant for Perry and Abecedarian were multiples of the levels of investment in present-day state preschool programs.} (Whitehurst 2013b)\textsuperscript{33}

\textit{[A] nationwide expansion of early education . . . won’t have the highly motivated administrators and hand-picked staffs that demonstration projects enjoy.} (Murray 2013)\textsuperscript{34}

\textit{The circumstances of the very poor families of the Black children who were served by these model programs 30 to 40 years ago are very different from those faced by the families that are presently served by publicly funded preschool programs. . . . [Forty] years ago other government supports for low-income families were at much lower levels and pre-K was not widely available for anyone.} (Whitehurst 2013b)\textsuperscript{35}

However, the research evidence for early childhood education encompasses more than Perry and Abecedarian. Evidence comes from more recent studies of large-scale programs with more modest costs. As discussed in Chapter 2, many large-scale state and local pre-K programs have strong effects on short-run test scores; such effects predict sizable long-run earnings benefits. The Chicago Child-Parent Center studies provide direct evidence of long-run benefits. CPC was a large-scale program, with similar costs per child to many state and local pre-K programs.

More recent programs have smaller earnings benefits than Perry and Abecedarian. CPC has average earnings effects of 8 percent, and many state and local pre-K programs have effects from 6 to 10 percent, whereas Perry had earnings effects of 19 percent and Abecedarian had earnings effects of 26 percent. But because costs are lower, the benefit-cost ratio for recent pre-K programs is still strong. As Chapter 3 showed, a full-day pre-K program might have a ratio of earnings benefits to costs of over 5-to-1, whereas Abecedarian’s high costs lead to a lower ratio of earnings benefits to costs of 1.5-to-1.
Today’s pre-K programs are run at a large scale by state agencies and local public schools, not by expert researchers who handpick a few great teachers. For example, the public schools in Tulsa, Boston, and Chicago all run successful pre-K programs.

The presence of competing programs complicates benefit-cost analyses. If a new pre-K program substitutes for existing programs, this reduces the new program’s net earnings benefits. But net costs of the new program are also reduced, because it reduces spending on existing programs. The ratio of benefits to costs need not decline.

Fewer of these recent evaluations are random assignment experiments. Random assignment experiments are expensive, and they are difficult to set up for a large-scale program. Such experiments will always be rare and will tend to be small-scale. But, as argued in Chapter 2, these recent evaluations are reliable because they have good comparison groups, chosen by natural experiments.

The recent studies look at test score effects in the short term, not directly at adult earnings effects. Benefits for adult earnings are inferred from test score effects. But this is inevitable if one wants timely evaluations of recent programs.

Finally, the Perry and Abecedarian evidence is still relevant to current policy issues. Today’s Educare program is similar to Abecedarian. Today’s pre-K programs have many similarities to Perry. Today’s programs do have larger class sizes than Perry, and they usually only last one year rather than Perry’s two years. On the other hand, Perry was a half-day program, whereas many current pre-K programs are full-day. All in all, Abecedarian and Perry are similar enough to today’s programs to provide useful information for current policy.

CRITICISM 2: Early childhood education is argued to be ineffective because the recent Head Start experiment has test score effects that quickly faded to statistical insignificance.

Critics have argued that the recent random assignment experiment on Head Start trumps other research evidence and shows that large-scale early childhood education is ineffective:
The Head Start Impact Study is a randomized controlled trial, the gold-standard for evaluating the effectiveness of social and health programs. . . . The findings . . . are that there were effects favoring Head Start children on some outcome variables at the end of the Head Start year. However, these impacts did not persist. Both in the kindergarten and first grade follow-up data . . . , and the third grade follow-up data . . . , there were no reliable differences in outcomes for children who won the lottery to attend Head Start vs. those who lost that lottery and served as the control group. (Whitehurst 2013a)

Head Start’s impact is no better than random. (Wall Street Journal, editorial published on February 27, 2013).36

These criticisms are based on the Head Start experiment’s finding that estimated effects decline by third grade so that they are insignificantly different from zero. Based on literacy and math tests administered both at the end of Head Start and the end of third grade, the effects of Head Start decline by over 70 percent. The point estimate of test score effects at the end of third grade still predicts that Head Start will increase lifetime earnings by 1.2 percent, which is a lot of money over a career. But because of statistical uncertainty, we cannot reject that the true effect is zero or negative. The true effect could also be three times as great. However, if one begins with a skeptical attitude that the true effect is zero until proven otherwise, then the Head Start results at Grade Three do not overcome that skepticism.37 If the Head Start random assignment experiment was the only evidence on the effects of early childhood education, skepticism about these programs’ effectiveness would be warranted.

However, many early childhood programs have fading test score impacts, but still significantly improve adult outcomes. As shown in Chapter 2, test score fading occurs not only in Head Start, but in many early childhood programs, including Perry Preschool, the Abecedarian Project, and the Chicago Child-Parent Center Program. Despite this test score fading, adult earnings effects reemerge. The initial test score effects are better predictions of adult earnings effects than the
faded test score effects. Figure 4.1 illustrates the evidence for these four programs (Head Start, Perry, Abecedarian, CPC).38

This fading and reemergence of effects could be due to non-cognitive skills, which are important to adult earnings but harder to measure using standardized tests. Social skills and character skills are at least as important as cognitive skills in making a worker more employable and more productive. Worker employability and productivity will depend upon social skills such as how a worker relates to supervisors, coworkers, and customers, and upon character skills such as reliability in showing up at work on time and being persistent in finishing work assignments. Cognitive skills also matter to adult earnings, but these cognitive skills must be applied effectively, which depends on character skills and social skills.

Figure 4.1 Predicted Percentage Effects on Adult Earnings of Early Childhood Programs, Based on Test Scores versus Adult Outcomes

SOURCE: Author’s calculations, based on research described in text and notes.
Pre-K can get children off to a good start by developing basic cognitive skills, social skills, and character skills. Over time, these skills build on themselves. If a child has better cognitive skills, social skills, and character skills at kindergarten entrance, that child will learn more in kindergarten and will develop still more self-confidence, ability to learn, and ability to operate socially in school. And so on in first grade and higher grades—at each grade, the skills built in previous grades lead to additional learning. By third grade, the boost to skills provided by pre-K is inadequately measured by standardized tests, as much of the boost comes from social and character skills that are harder to measure. But the initial boost to cognitive skills at kindergarten entrance is a catalyst that leads to these broader skills, by increasing self-confidence and the ability to learn. The important catalytic role of cognitive skills is why the end-of-program test score gain helps predict adult earnings.

Another important point is that the recent Head Start experiment may not represent the effectiveness of the best pre-K programs, Head Start in the past, or Head Start compared to no pre-K. Several studies show that initial test score effects of Head Start are one-half to two-thirds of some other pre-K programs. Furthermore, the recent Head Start experiment’s results are unusual in the rapidity of fade-out, compared both to other pre-K programs and to Head Start in the past. For example, Deming (2009) found little fading of Head Start’s initial effects in elementary school, with more fading occurring in middle school.

A possible explanation is that the Head Start experiment does not compare Head Start with no pre-K; it compares a treatment group offered enrollment in Head Start with a control group that was supposed to be denied admission to Head Start but often attended some pre-K program. Only 80 percent of the treatment group enrolled in Head Start. About half of the control group attended some pre-K program, including 14 percent in Head Start and 35 percent in some other pre-K program (Puma et al. 2012, p. xix; U.S. Department of Health and Human Services 2010). If these other pre-K programs were more
effective than Head Start, the experiment’s net impact would be reduced.

In recent years, Head Start has faced more competition. Studies showing long-run benefits of Head Start are necessarily examining Head Start in the past, when low-income families had fewer high-quality alternatives. In recent years, increased availability of state pre-K means that Head Start needs to up its quality to justify its higher costs. But today’s lower net Head Start impacts need not imply that high-quality pre-K makes no difference compared to no pre-K.

Recent Head Start reforms may have increased quality, compared to Head Start at the time of the experiment (2002–2003). From 2003 to 2006, more literacy instruction was pushed in Head Start. Research suggests that Head Start children gained more in literacy in 2006 and 2009 than was true in 2003 (Barnett 2013).

CRITICISM 3: Why haven’t Oklahoma and Georgia, two states that have prominently adopted universal pre-K, been more obviously successful in improving test scores and other social indicators?

Pre-K advocates have sometimes pointed to Oklahoma and Georgia as models (Lerner 2012; Obama 2013). These states were early adopters of broader pre-K access (Georgia in 1995, Oklahoma in 1998), and have moved a long way toward universal access. (Oklahoma has 74 percent of all four-year-olds in its state program; Georgia has 58 percent [Barnett et al. 2013].) Oklahoma’s program meets high-quality standards (Barnett et al.) and has received favorable evaluations (Wong et al. 2008). (Georgia’s pre-K program has more mixed reviews.)

But, critics argue, if broad access to pre-K is so great, why haven’t Oklahoma and Georgia made more progress? Shikha Dalmia and Lisa Snell of the Reason Foundation hold that “neither state program has demonstrated major social benefits. . . . The average NAEP reading score for Oklahoma fourth-graders dropped four points between 1998
36 Bartik

and 2011—although it went up nine points for Georgia. . . . Oklahoma remains below the national average and Georgia has just reached the national average” (Dalmia and Snell 2013).

The Wall Street Journal editorial board quotes President Obama, from a Georgia speech promoting federal support for preschool, as saying that “education has to start at the earliest possible age. . . . If you are looking for a good bang for your educational buck, this is it right here.” The Journal goes on to assert that “Mr. Obama is right that the state is a good example of what universal pre-K can buy. Georgia’s fourth- and eighth-grade reading, math and science scores all trail the national average” (Wall Street Journal 2013).

Both pre-K critics and advocates need to recognize that analysis of test scores in just one or two states is subject to great uncertainty. Because average state test scores are frequently buffeted by demographic or economic changes, detecting the influence of any policy on test scores in one or two states is statistically difficult. Even if a state’s pre-K program improves test scores enough to predict large long-term benefits, these improvements can easily be masked by other influences.

Part of the empirical challenge is that even slight improvements in test scores are sufficient to predict large economic gains. A two percentile gain in average elementary school test scores—moving from the 50th to the 52nd percentile on a standardized test—is enough to predict that future adult earnings will increase by 1 percent (Chetty et al. 2011). For the average American worker, career earnings exceed $1.5 million.41 A 1 percent boost would increase career earnings by over $15,000, a large benefit. Demographic and economic changes can easily affect test scores by more than two percentiles. Because of this statistical “noise” in a state’s average test scores, it is hard for an examination of one state’s test scores to rule out test score increases that would be meaningful economically.42

A good recent study of pre-K in Oklahoma and Georgia tries to minimize this uncertainty by combining many years of data (Cascio and Schanzenbach 2013). This study estimates average state test
score effects that predict that benefits exceed costs by at least 3-to-1. However, because of the statistical noise inherent in focusing on test score effects in just two states, these estimated test score effects are not statistically significantly different from zero at conventional levels of significance used by researchers.

People love anecdotes. We are tempted to believe that the truth can be discovered by the examples of one or two states. But this is difficult to do. If states are laboratories of democracy, this lab gets much more precise results with many “test subjects” (states), rather than with just one or two states as test subjects. An alternative way to get the larger numbers needed for statistical precision is to compare many individuals, some of whom participated in early childhood education versus similar nonparticipants.

**CRITICISM 4: The recent Tennessee pre-K experiment shows that test score effects quickly fade, which raises questions about whether state pre-K programs are effective.**

A recent random-assignment experimental study of Tennessee’s pre-K program found some effects at the end of pre-K, but effects faded by the end of kindergarten or the end of first grade (Lipsey et al. 2013a, b). According to one critic,

*I see these findings as devastating for advocates of the expansion of state pre-K programs. This is the first large scale randomized trial of a present-day state pre-K program. Its methodology soundly trumps the quasi-experimental approaches that have heretofore been the only source of data on which to infer the impact of these programs. . . . The most defensible conclusion is that these statewide programs are not working to meaningfully increase the academic achievement or social/emotional skills and dispositions of children from low-income families.*

(Whitehurst 2013c)

One problem with the Tennessee study is its considerable sample attrition, which may lead to bias. In the first cohort of children, the
Tennessee study only obtained parental consent to have tests administered to the children for 46 percent of the pre-K participants and for 32 percent of the control group. For the sample with test data, the treatment and control groups may differ greatly in unobserved characteristics, particularly since the attrition differs so much in the two groups. Any test score differences between the two groups will be due to some combination of true effects of the pre-K program and bias due to unobserved characteristics. The Tennessee researchers do a good job of minimizing biases by controlling for the child’s and family’s observed characteristics. But such controls cannot adjust for biases due to unobserved characteristics.45

Because of these problems, the Tennessee study does not trump other studies of state and local pre-K programs. As discussed in Chapter 2, these other studies try to estimate effects of pre-K by natural experiments in which pre-K access varies because of where a family lives or a child’s age. The argument is that such natural variation in pre-K access will cause the treatment and comparison groups to be similar in unobserved characteristics. While a random assignment study with few problems may be superior to natural experiments, a random assignment study with sizable attrition cannot claim such superiority.

Another important point is that Tennessee’s pre-K program is not as high-quality as other state and local pre-K programs. For example, Tennessee spends a little under $6,000 per child annually for a full-day pre-K program. According to Barnett et al. (2013), Tennessee would have to spend at least $2,000 extra per child to consistently deliver quality.46 In contrast, pre-K programs such as those in Tulsa, Chicago, and Boston all pay amounts per child that equal or exceed guidelines for the funding necessary to reach quality standards. Even if Tennessee’s program is ineffective, this need not imply the ineffectiveness of state and local pre-K programs of better quality.

One study of one state’s program rarely trumps all other studies. Any single study has limitations, which is why policy should be based
Criticisms of the Research Evidence 39

on many studies. The case for early childhood education is that high
quality programs work, not that all programs work.

CONCLUDING COMMENTS

A persistent theme in these criticisms is that the effectiveness of early childhood education is uncertain. Uncertainty argues for postponing action until we know more. Russ Whitehurst of Brookings makes precisely this argument in discussing the Tennessee results: “Maybe we should figure out how to deliver effective programs before the federal government funds preschool for all” (Whitehurst 2013c). But, as this book will discuss in Chapter 5, we do know something about how to deliver effective programs.