SCHOOL of PUBLIC POLICY

12
Five-State Pre-K Evaluation

Douglas J. Besharov
Peter Germanis
Caeli A. Higney
and
Douglas M. Call

September 2011

Maryland School of Public Policy
Welfare Reform Academy
www.welfareacademy.org

Part of a forthcoming volume
Assessments of Twenty-Six Early Childhood Evaluations
by Douglas J. Besharov, Peter Germans, Caeli A. Higney, and Douglas M. Call

Note: This report is open to public comments, subject to review by the forum moderator. To leave a comment, please send an email to welfareacademy@umd.edu or fill out the comment form at http://www.welfareacademy.org/pubs/early_education/chapter12.html.

12
Five-State Pre-K Evaluation

The pre-K programs of Michigan, New Jersey, Oklahoma, South Carolina, and West Virginia are state-funded, school-based programs focused exclusively on providing education to three- and four-year-olds and are considered to be among the strongest pre-K programs in the nation.

Steve Barnett and his colleagues at the National Institute of Early Education Research and Northwestern University ("the NIEER team") conducted an evaluation of pre-kindergarten programs in Michigan, New Jersey, Oklahoma, South Carolina and West Virginia. Three of the states limited pre-K enrollment to disadvantaged children (Michigan, South Carolina) or to children residing in counties with a high percentage of disadvantaged children (New Jersey). The other two (Oklahoma and West Virginia) did not have any limitation to enrollment based on the children's status. The NIEER team used a regression-discontinuity design (RDD) to compare test scores of sample children who had attended pre-K and were entering kindergarten in September 2004 with the test scores of sample children who were entering pre-K in September 2004. Using two different analyses, the NIEER team reports that the pre-K programs in the five states produced, on average, medium positive effects on the children's reading test scores and small positive effects on the children's vocabulary and math scores.[1]

However, the pooled analysis may have used an inaccurate functional form of the data, generating potentially biased estimates, and the findings for many of the effect sizes for the instrumental variable (IV) individual state analysis are not statistically significant. Moreover, these findings are of questionable generalizability. First, although the overall effects are statistically significant, many of the effects for the individual states are not. Second, a number of individual schools and also some school districts in the five states did not consent to be part of the evaluation which raises questions about the generalizability of the findings within each state. Finally, the five states included in the study were considered to have the highest quality pre-K programs and, therefore, not representative of the rest of the country (although they would demonstrate the potential for such programs).

---

[1]See Vivian C. Wong, Thomas D. Cook, W. Steven Barnett, and Kwanghee Jung, "An Effectiveness-Based Evaluation of Five State Pre-Kindergarten Programs," *Journal of Policy Analysis and Management* 27, no. 1 (2008); and W. Steven Barnett, Kwanghee Jung, Vivian Wong, Tom Cook, and Cynthia Lamy, *Effects of Five State Prekindergarten Programs on Early Learning* (New Brunswick, NJ: National Institute of Early Education Research, October 2007).

Program Design

Program group. The target program group differs by state, based on income, age, and location factors.

Income. Michigan and South Carolina both target children from at-risk families (as defined by the state), where income below a certain threshold qualifies a child as being at-risk. New Jersey has three different pre-K programs; only the largest of these programs, the Abbott program, was included in this evaluation. The Abbott program operates in thirty-one New Jersey school districts where, at the time of the inception of the program, 40 percent of the children qualified for subsidized school lunch.[2] In these school districts, pre-K is available for all children regardless of income. (The pre-K programs in these districts are called Abbott programs.) Enrollment in Oklahoma and West Virginia is not limited by income.

Age. Michigan, Oklahoma, and South Carolina limit state-sponsored pre-K to four-year olds. In the New Jersey Abbott program, all three-year olds are also eligible for pre-K. West Virginia also serves some three year olds, but only a very small percentage of the eligible population.

Location. As mentioned above, in New Jersey, children who live in Abbott program districts are eligible for pre-K. Although New Jersey has two other pre-K programs for children who do not live in these school districts, the Abbott program is the most well-funded and serves the vast majority of children in pre-K in the state. In West Virginia, the local school district decides if enrollment is based on a lottery or on a first-come-first serve basis until funds are exhausted. Oklahoma allows each school district the option of offering pre-K; 91 percent of the school districts have opted to do so. Both Michigan and South Carolina's pre-K enrollment policies are unaffected by location.

Overall, Oklahoma is the most "universal" of the programs, serving 65 percent of eligible four-year olds; Michigan and New Jersey's Abbott program were the least universal, serving only 19 percent of all four-year olds in their respective states. For the total sample in all five states, 48 percent of pre-K enrollees were white, 24 percent were black, and 23 percent were Hispanic. The remaining 7 percent were other races. Fifty-five percent of the children were eligible for free or reduced lunches. The comparison group had a greater percentage of black children and a smaller percentage of Hispanic children than the program group, but the differences were not statistically significant.

Services. Unlike other early childhood interventions, the five state pre-K programs do not

[2]W. Steven Barnett, National Institute of Early Education Research, e-mail to Douglas Call, March 7, 2008.

provide non-classroom services to either parents or children, focusing almost exclusively on educational provision. New Jersey's Abbott program is the exception, providing additional child care services in pre-K classrooms. While the pre-K program operates each weekday for six hours throughout the entire school year, the state-funded additional child care services are available up to ten hours a day, five days a week, year round. In contrast, South Carolina's pre-K operates only 2.5 hours a day, five days a week. Michigan, Oklahoma, and West Virginia each allowed the districts to choose whether to operate a part-day or full-day pre-K program.

In all of the five states, children are served in either public schools, Head Start centers, or other private child care facilities. Oklahoma, South Carolina, and West Virginia require adult-child ratios of 1:10, and Michigan and New Jersey require adult-child ratios of 1:8. Maximum class sizes varied between fifteen in New Jersey, eighteen in Michigan, and twenty in Oklahoma, South Carolina, and West Virginia. West Virginia was the only state that did not require that teachers have a bachelor's degree to teach pre-K.

The Evaluation. The NIEER team used a regression-discontinuity design to evaluate the pre-K programs of the five states. This type of evaluation is a non-experimental design that determines placement in the program and control group based on a cutoff score or a selection variable.[3] Battistin and Rettore of the Centre for Economic Policy Research write, "By exploiting the fact that the subjects assigned to the comparison and the intervention group solely differ with respect to the variable on which the assignment to the intervention is established, one can control for the confounding factors just by contrasting marginal participants to marginal non-participants . . . the term *marginal* refers to those units *not too far* from the threshold for selection."[4] Estimated program effects are derived through regression analyses using data of participants near the cutoff.[5] Regression-discontinuity designs also need sample sizes that are much larger than random assignment evaluations to achieve the same statistical precision.[6]

The NIEER team's regression-discontinuity design made use of the strict birthday cutoffs in the states—children with birthdays prior to the cutoffs were allowed to enroll in pre-K in September 2003 while children with birthdays on or after the cutoffs were not allowed to enroll in pre-K until September 2004. This design attempts to control for selection bias by comparing the

---

[3]Peter H. Rossi, Howard E. Freeman, and Mark W. Lipsy, *Evaluation: A Systematic Approach*, 6th ed. (Thousand Oaks, CA, 1999).

[4]Erich Battistin and Enrico Rettore, *Another Look at the Regression-Discontinuity Design* (London: Centre for Economic Policy Research, February 2002), 3.

[5]Rossi, Freeman, and Lipsy, 1999.

[6]William M. K. Trochim, *The Regression-Discontinuity Design: An Introduction* Research Methods Paper Series (Chicago, IL: Thresholds National Research and Training Center on Rehabilitation and Mental Illness, 1994).

test scores of children near the cutoffs on either side. The NIEER team writes that the design "can be viewed as modeling the relationship between the assignment variable (age) and the relationship prior to treatment. The pre-cutoff sample models the relationship prior to treatment. The post-cutoff sample is used to model the relationship after the treatment."[7]

In September 2004, the NIEER team administered the Peabody Picture Vocabulary Test (PPVT), the Applied Problems sub-test of the Woodcock-Johnson Achievement Test (WJ), and the Print Awareness sub-test of the Preschool Comprehensive Test of Phonological & Print Processing (Pre-CTOPPP) to a total of 2,728 kindergarten students and 2,550 four-year-old pre-K students in Michigan, New Jersey, South Carolina, Oklahoma, and West Virginia.[8] The children in kindergarten who were tested consisted of children who had been enrolled in pre-K the previous year.[9] After determining there were no statistically significant differences in the demographics of the two groups, the NIEER team then constructed the regression-discontinuity design using the test scores of the children entering kindergarten as the program group and the test scores of the children entering pre-K as the comparison group.

The NIEER team evaluated the state data separately (individual state sample) and pooled into one larger sample with interaction variables to obtain state estimates (pooled sample). For each sample, the NIEER team conducted two analyses: a "sharp" analysis that did not include "fuzzy" cases (cases where the child's birthday was inconsistent with the cutoff requirement of the program) and an instrumental variable analysis which included "fuzzy" cases. Ultimately, the NIEER team presented the results from the sharp analysis for the pooled sample, and the IV analysis for the individual state sample. For the latter sample, the NIEER team ran separate regressions for each state and then averaged the results to determine the overall result. Each analysis has different strengths and weaknesses. The sharp pooled analysis has smaller standard errors and is statistically more powerful than the IV separate state analysis because of its large sample size, but might provide inaccurate estimates if the specified functional form is inaccurate. The IV individual state analysis can provide a better fit for the functional forms of the observations, providing more accurate estimates, but this risks "overfitting the data and

[7]W. Steven Barnett, Kwanghee Jung, Vivian Wong, Tom Cook, and Cynthia Lamy, *Effects of Five State Prekindergarten Programs on Early Learning* (New Brunswick, NJ: National Institute of Early Education Research, October 2007), 8.

[8]Vivian C. Wong, Thomas D. Cook, W. Steven Barnett, and Kwanghee Jung, "An Effectiveness-Based Evaluation of Five State Pre-Kindergarten Programs," *Journal of Policy Analysis and Management* 27, no. 1 (2008): 131.

[9]W. Steven Barnett, National Institute of Early Education Research, e-mail message to Douglas Call, March 7, 2008.

suffer[ing] from reduced power."[10]

In the sharp pooled analysis, the NIEER team determined the functional form was roughly linear for the PPVT and the WJ which allowed them to include all children in the data set.[11] In the IV separate state analysis, the NIEER team looked at the functional form of each of the different tests in each state, concluding that the linear functional form "fit best for 9, the cubic for 3, and the quadratic for 2."[12] Because of the differing functional forms, the NIEER team examined the six-month estimates to ensure the proper functional form, but ultimately used the twelve-month estimates.

Major Findings.

Among four-year-olds in the five state programs, those who were enrolled in pre-K showed positive effects for all three types of test scores. Using Jacob Cohen's widely accepted guidelines for effect sizes (small=0.20 SD, medium=0.50 SD, and large=0.80 SD; see Appendix 1 for a further discussion of effect sizes and their interpretation),[13] no results can be considered large. The effect sizes of the Pre-CTOPPP in both the pooled and IV analysis, and the WJ in the sharp pooled analysis can be considered medium effects. The PPVT test scores in both analyses cannot even be considered to be small gains. In addition, there was large variance between the states for each score and no one state had consistently higher effect sizes on test scores than the other states.

Cognitive. In the sharp pooled analysis, the NIEER team found the largest average effect size on the Pre-CTOPPP (0.74 SD), followed by the WJ (0.43 SD) and the PPVT (0.18 SD). The effect sizes in the IV individual state analysis were smaller than in the sharp pooled analysis. The largest average effect size was on the Pre-CTOPPP (0.70 SD), the next largest was on the WJ (0.29 SD), and the smallest was on the PPVT (0.14 SD). For both analyses, findings varied widely between states and in some states, the findings were not statistically significant. (See Table 1.)

---

[10]W. Steven Barnett, National Institute of Early Education Research, e-mail message to Douglas Call, March 7, 2008.

[11]"When the response functions are parallel and linear, one can generalize treatment effects across the entire distribution of the assignment variables." W. Steven Barnett, Kwanghee Jung, Vivian Wong, Tom Cook, and Cynthia Lamy, *Effects of Five State Prekindergarten Programs on Early Learning* (New Brunswick, NJ: National Institute of Early Education Research, 2006), 16.

[12]W. Steven Barnett, Kwanghee Jung, Vivian Wong, Tom Cook, and Cynthia Lamy, *Effects of Five State Prekindergarten Programs on Early Learning* (New Brunswick, NJ: National Institute of Early Education Research, 2006), 16.

[13]Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Hillsdale, NJ: Lawrence Erlbaum, 1988), 535.

Table 1. Effect Sizes for Cognitive Findings						
	Vocabulary (PPVT)		Math (WJ)		Reading (Pre-CTOPPP)	
	Pooled	IV	Pooled	IV	Pooled	IV
Michigan	—	—	0.51	0.47	0.78	0.96
New Jersey	0.34	0.36	0.19	0.23	0.46	0.50
Oklahoma	0.32	0.29	0.49	—	0.54	—
South Carolina	—	—	NA	NA	0.81	0.79
West Virginia	—	—	0.52	—	1.10	—
Total average	0.18	0.14	0.43	0.29	0.74	0.70

Although data related to race and income was collected, no effect sizes were reported for these variables.

School readiness/performance. Data apparently either not collected or not reported.

Socioemotional development. Data apparently either not collected or not reported.

Health. Data apparently either not collected or not reported.

Behavior. Data apparently either not collected or not reported.

Crime/delinquency. Data apparently either not collected or not reported.

Early/nonmarital births. Data apparently either not collected or not reported.

Economic outcomes. Data apparently either not collected or not reported.

Effects on parents. Data apparently either not collected or not reported.

Benefit-cost findings. The NIEER team estimates that the total cost (including all local, state, and federal costs) was about $5,300 per student in Michigan, about $11,000 per student in New Jersey, about $6,500 per student in Oklahoma, about $3,400 per student in South Carolina,

and about $7,250 per student in West Virginia.[14] All dollar amounts are in 2007 dollars.

Overall Assessment

Program theory. The five state pre-K programs are based on the theory that early education programs can build children's human capital by engaging in "age-appropriate cognitive tasks."[15]

Program implementation. The evaluation does not report on the program implementation in the five states.

Assessing the randomization. The NIEER team did not use random assignment in their evaluation.

Assessing statistical controls in experimental and nonexperimental evaluations. In general, using a birthday cutoff in a regression-discontinuity design may introduce selection bias as parents decide when their children will enter pre-K. Michael Puma, president of Chesapeake Research Associates LLC, explains that parents with "more able" children who have birthdays near the cutoff may enroll their children in pre-K while parents with "less able" children may wait another year to enroll them in pre-K. When comparing the program group and the comparison group near the cutoff, the program group would then consist of "more able" children than the comparison group.[16]

A major concern with the reported findings of the sharp pooled analysis is that the functional form of the data was incorrectly specified. As mentioned above, regression-discontinuity designs only maintain the properties of a random assignment design if the functional form is correct and the necessary adjusting variables are added to the regression. Barnett writes that if the correct models are not all linear then "the pooled analysis may raise the risk of missing real difference in the correct functional form and, thereby, produc[e] biased estimates for some

[14]Vivian C. Wong, Thomas D. Cook, W. Steven Barnett, and Kwanghee Jung, "An Effectiveness-Based Evaluation of Five State Pre-Kindergarten Programs," *Journal of Policy Analysis and Management* 27, no.1 (2008): 128-129.

[15]Vivian C. Wong, Thomas D. Cook, W. Steven Barnett, and Kwanghee Jung, "An Effectiveness-Based Evaluation of Five State Pre-Kindergarten Programs," *Journal of Policy Analysis and Management* 27, no.1 (2008): 122.

[16]Michael Puma, Chesapeake Research Associates LLC, e-mail message to Douglas Besharov, December 27, 2005.

states."[17] When specifying the proper functional forms of the tests, the NIEER team identified that on the PPVT, Michigan's data were quadratic while the remaining four states were linear. On the WJ, West Virginia's data were quadratic, Oklahoma's were cubic, and the remaining two states (South Carolina did not report WJ scores) were linear. In the sharp pooled analysis, both of the regressions assumed a linear functional form. For the Pre-CTOPPP, New Jersey and Oklahoma's data were cubic while the remaining three states were linear, for which the regression assumed a quadratic functional form. This possible misspecification of the underlying functional form raises questions about the viability of the pooled analysis's findings.

In the IV individual state analysis, only slightly more than half of the calculated effect sizes for test scores across all states were statistically significant and only the Pre-CTOPPP was statistically significant in the majority of the states. Thus, on the PPVT and the WJ, the reported average effect sizes are based on a combination of statistically significant and non-statistically significant findings.

There were also some problems associated with the initial sampling of the students. The NIEER team randomly sampled classrooms and children to obtain their sample in four of the five states. In New Jersey, a stratified random sample was used in the state's largest pre-K program. However, 41 percent of children in West Virginia opted not to participate and the Detroit School District granted permission too late to be included in the study. Some of the children who opted not to participate were non-randomly replaced in the sample while others were not replaced. The NIEER team did not mention how many students were non-randomly replaced, but they do mention that the sample did not "perfectly represent all students enrolled in a state's pre-K program."[18] This method of replacing children in the sample raises questions of selection bias, as the children who were selected might be better students than those who opted not to participate.

Finally, the NIEER team reported that the Oklahoma estimates are less certain than the other state estimates because of the lack of observations around the cutoff. Although an overall lack of observations around the cutoff prevented the NIEER team from running regressions using data from within three months of the cutoff, Oklahoma has even fewer cases near the cutoff making it difficult to judge the functional form of the data and making the estimates less certain.[19]

---

[17]W. Steven Barnett, Kwanghee Jung, Vivian Wong, Tom Cook, and Cynthia Lamy, *Effects of Five State Prekindergarten Programs on Early Learning* (New Brunswick, NJ: National Institute of Early Education Research, 2006), 16

[18]Vivian C. Wong, Thomas D. Cook, W. Steven Barnett, and Kwanghee Jung, "An Effectiveness-Based Evaluation of Five State Pre-Kindergarten Programs," *Journal of Policy Analysis and Management* 27, no.1 (2008): 127–128.

[19]Vivian C. Wong, Thomas D. Cook, W. Steven Barnett, and Kwanghee Jung, "An Effectiveness-Based Evaluation of Five State Pre-Kindergarten Programs," *Journal of Policy Analysis and Management* 27, no.1 (2008): 146.

Sample size. The initial sample size and the sample size that was used for the individual state IV analyses was 5,278. New Jersey was the state with the largest sample with 2,072. Michigan had 871, Oklahoma 838, South Carolina 777 and West Virginia 720. The sharp pooled data analysis used a sample of 5,071 children from five different states after dropping 207 children with birthdays that did not match with the cutoff dates.

The NIEER team expresses concern with the sample size for the IV individual state analysis because a separate regression was run for each state. Smaller sample sizes in each state may have led to the large variance in the final state effect sizes. The NIEER team mentions that if larger sample sizes were used in future studies "some of the apparent variation might be reduced."[20]

In addition, there is a potential problem of non-response bias. The NIEER team originally targeted 7,600 children for the evaluation. As mentioned above, only about 70 percent of children targeted for participation in the group completed the evaluation. As the characteristics of the children that did not participate are unknown, it is possible that the final estimates could be biased if the decision not to participate was non-random.

Attrition. Because both groups (program and control) were tested only once, attrition is not a factor.

Data collection. The data collection relied on the PPVT, the WJ, and the Pre-CTOPPP, all of which are nationally recognized assessment tools and on school records.

Measurement issues. The evaluation relies on standard cognitive and achievement tests.

Generalizability. The NIEER team report that these five states are among the top pre-K programs in the country based on quality standards. They write, "As encouraging as these results are, it is difficult extrapolating from them to the nation at large."[21]

Replication. As of publication, thirty-eight states have developed pre-K programs.

Evaluator's description of findings. The NIEER team is quite positive about their findings. They write that "the results of the study add to the evidence that high quality public

[20]W. Steven Barnett, Kwanghee Jung, Vivian Wong, Tom Cook, and Cynthia Lamy, *Effects of Five State Prekindergarten Programs on Early Learning* (New Brunswick, NJ: National Institute for Early Education Research, October 2007), 26.

[21]Vivian C. Wong, Thomas D. Cook, W. Steven Barnett, and Kwanghee Jung, *An Effectiveness-Based Evaluation of Five State Pre-Kindergarten Programs Using Regression-Discontinuity* (New Brunswick, NJ: National Institute for Early Education Research, June 2007, 33.

preschool education can improve learning and development on a large scale for both targeted and general populations"[22] and that "our study adds to the evidence that high-quality public preschool education programs can be sound investments from a purely economic perspective." However, they do warn that "the five states have better than average pre-K programs, effects were stronger for alphabet learning than for more general pre-reading or mathematical skills, and long-term effects cannot be ascertained yet."[23]

Their overall conclusions, however, seem to be at odds with the description of their findings. They admit there is no definitive reason why the findings widely varied between states and admit it might be due to sampling error,[24] and that this is a particular problem with the IV individual state estimates.[25] In addition, many of the overall effect sizes are either small (according to the Cohen scale) or not statistically significant.

Evaluator's independence. Steve Barnett and Kwanghee Jung are from the National Institute for Early Education Research at Rutgers University. Vivian Wong and Thomas Cook are professors at Northwestern University and Cynthia Lamy is from the Robin Hood Foundation. None of the researchers are affiliated with the state programs.

Statistical significance/confidence intervals. Statistical significant is measured and reported at the 1 percent, 5 percent, and 10 percent levels.

Effect sizes. Effect sizes were calculated as the "estimated regression coefficient divided by the standard deviation of the control group."[26] For all children, effect sizes ranged from 0.14 to 0.70 SD using the IV analysis and from 0.18 to 0.74 SD using the pooled analysis. The weighted effect sizes of the IV analysis ranged from 0.16 to 0.68 SD. Using the traditional demarcations for

---

[22]W. Steven Barnett, Kwanghee Jung, Vivian Wong, Tom Cook, and Cynthia Lamy, *Effects of Five State Prekindergarten Programs on Early Learning* (New Brunswick, NJ: National Institute of Early Education Research, 2006), 26.

[23]Thomas D. Cook and Vivian C. Wong, "The Warrant for Universal Pre-K: Can Several Thin Reeds make a Strong Policy Boat?" *Social Policy Report* vol. 21, no. 3 (2007): 14, http://www.srcd.org/documents/publications/spr/21-3_early_childhood_education.pdf (accessed October 16, 2008).

[24]W. Steven Barnett, Kwanghee Jung, Vivian Wong, Tom Cook, and Cynthia Lamy, *Effects of Five State Prekindergarten Programs on Early Learning* (New Brunswick, NJ: National Institute of Early Education Research, 2006), 21

[25]W. Steven Barnett, Kwanghee Jung, Vivian Wong, Tom Cook, and Cynthia Lamy, *Effects of Five State Prekindergarten Programs on Early Learning* (New Brunswick, NJ: National Institute of Early Education Research, 2006), 24

[26]W. Steven Barnett, National Institute of Early Education Research, e-mail message to Douglas Call, June 6, 2008.

measuring effect sizes, the effect sizes on the lower bound of the range can be considered not practically significant or meaningful. The effect sizes of the upper bound of the range can be considered medium effects.

One of the evaluators, Thomas Cook, warns that the effect sizes found in this evaluation are not comparable to effect sizes found in the Head Start evaluations due to the different population served by the programs, the focus of the pre-K programs on cognitive gains only, and the generalizability of the nationally representative Head Start evaluations compared to the local pre-K evaluation.[27]

Sustained effects. The evaluation did not examine post-intervention impacts.

Benefit-cost analysis. Apparently not performed.

Cost-effectiveness analysis. Apparently not performed.

[27]Thomas Cook, "Pre-K Programs: Which Ones Make a Difference?" (presentation, IPR Policy Briefing, Washington, DC, May 19, 2006), http://www.northwestern.edu/ipr/events/briefingmay06-cook/slide1.html (accessed November 10, 2008).

Commentary

W. Steven Barnett and Kwanghee Jung[1]

We appreciate the opportunity to respond to the review of our evaluation of five state pre-K programs using a regression discontinuity design (RDD) approach. Although the reviewers have made a good faith effort to describe and critique our study, they do not appear to have fully understood our approach. In our response, we try to correct some important misunderstandings about the sample and analytical methods. In addition, we explain why we disagree with the reviewers about the interpretation of our findings. Although it is wise to be cautious, the reviewers seem to us to be overly pessimistic in their assessment of our study. Caution does not dictate always adopting the most pessimistic assumption, nor does it justify concluding that any deviation from perfection, no matter how slight, has an overwhelming or even meaningful effect on a study's internal or external validity.

Sampling Issues

The state pre-K programs for which we estimated effects are, as a whole, above average in terms of their program standards and state commitment of funds. This suggests that our findings might over estimate the average effects of all state pre-K programs. We would expect programs with lower standards and less funding per child to have weaker results. However, it would be a mistake to exaggerate how different the programs we studied are from others, as the programs we studied encompass a broad range. They include high and low spenders. Local funding also tends to produce more equal funding levels than state spending per child alone suggests. Also, classroom experiences among state pre-K programs may vary less than one might suppose from differences in standards. Many other state pre-K programs are likely to be reasonably similar to those we studied in terms of the education that they provide to children.[1] Thus, our results might not generalize precisely, but in the context of the broader literature, they may still be reasonably interpreted as evidence that state pre-K programs generally have positive effects on learning.

---

[1]W. Steven Barnett is a professor at Rutgers University and the director of the National Institute for Early Education Research; Kwanghee Jung is a research professor at the National Institute for Early Education Research.

[1]Jennifer LoCasale-Crouch, Tim Konold, Robert Pianta, Carollee Howes, Margaret Burchinal, and Donna Bryant, "Observed Classroom Quality Profiles in State-Funded Pre-Kindergarten Programs and Associations with Teacher, Program and Classroom Characteristics," *Early Childhood Research Quarterly* vol.22, no.1: 3–17, http://www.sciencedirect.com/science?_ob=PdfDownloadURL&_uoikey=B6W4B-4K48JG9-1&_tockey=%23toc%236538%232007%23999779998%23644455%23FLA%23&_orig=search&_acct=C000049425&_version=1&_userid=961305&md5=a04e7315a639580a51c3be8d923a710b (accessed October 15, 2008).

All of the programs in our study have a primary focus on classroom services rather than some other service, such as home visiting, for example. This does not mean that they have no activities that extend beyond the classroom. For example, all of them make an effort to involve parents in the child's education. The level of effort devoted to this likely varies, but it can be a significant effort as in New Jersey where programs employ staff specifically dedicated to working with parents.

The composition of our sample varies from state to state partly because of variations in program eligibility and state demographics, but also because of differences in the research procedures that were agreed upon in each state. As the reviewers discuss, we had more difficulty obtaining consent from school districts and families in some states than in others. However, we did not have serious problems with refusals in every state. In some states, we had no refusals from programs and few parental refusals (because some states did not require parents to actively consent to child testing). Also, the reviewers incorrectly attribute the entire shortfall from planned sample sizes to refusals. Shortfalls also resulted from logistical difficulties including bad weather, data collector illness, and scheduling problems. Replacement of programs and children lost to the sample was random from among the participating programs and children. There is no evidence that estimated effects varied systematically between states that had relatively high and low rates of nonparticipation, and it is not at all certain how nonparticipation affects the results. The reviewers assume that the effect is to inflate estimates, but there is no evidence to support that assumption.

Design and Analysis Issues

As the reviewers note, the regression discontinuity approach has limitations that include: the need for a larger sample than a randomized trial for the same level of statistical power; the need to correctly specify the functional form in order to obtain unbiased estimates; and, adherence to the use of the birth date cutoff in determining program eligibility. It is because of these limitations that we conducted a wide range of analyses designed to test various assumptions, identify the correct functional forms, and assess the sensitivity of the results to alternative assumptions and specifications of model. We did not assume that we could necessarily identify one best model and so presented results from multiple approaches.

Unfortunately, the reviewers do not appear to fully understand the RDD approach or our procedures in applying that approach. For example, they appear to think that only data around the birth date cutoff are employed in the analysis. We used all of the data, though we also conducted analyses using only more narrow windows around the cutoff to see how that might have affected the results. The reviewers state that linearity is required to use all of the cases rather than just those around the cutoff, but this is not true. They appear to be especially confused about the "fuzzy," or instrumental variables (IV), analyses. Contrary to their assertions: the IV analyses do not remove the fuzzy cases (those that violate the assignment rule); IV analysis does not improve the chances of identifying the correct functional form; and, IV analysis increases (not decreases) the sample size. Nevertheless, IV analysis can result in larger standard errors.

One source of confusion appears to be that the reviewers incorrectly confounded the question of whether to pool the data across the 5 states to analyze a single model or to estimate a separate model for each state with the question of whether to conduct a sharp analysis discarding cases that violate the assignment rule or to conduct an IV analysis that includes all cases. We conducted both sharp and IV analyses on the pooled data, and we conducted both types of analyses on each individual state sample. The primary source of the differences between pooled analysis results and individual state analysis results is due to the pooling, and not to whether we employed a sharp analysis or an IV analysis that includes "fuzzy" cases. We present only the sharp results for the pooled sample, because the number of excluded cases relative to the total pooled sample size is so small that there is little difference in results. The IV analyses matter more for the individual state results, because the fuzzy cases tend to be concentrated in some states. Thus, we focus on the IV results in reporting the individual state analyses. The loss of statistical power when each state sample is analyzed in isolation is the primary reason that there are fewer statistically significant results in those analyses.

Thus, whether or not one prefers the IV approach to analysis is a different question, and of less importance, than whether one prefers a pooled analysis or separate analyses for each state. The pooled analyses maximize the sample size and, therefore, statistical power, which allows us to be more precise about effect sizes. In the pooled analyses, a given effect size is more likely to be identified as statistically significant, and greater power also contributes to our ability to identify the correct functional form for the regression equation. Alternatively, separate analyses for each state offer the most flexible approach to identifying the correct functional form as it allows all of the estimated relationships to vary completely from one state sample to another. However, this also has a downside, because analyzing each state in isolation ignores the information from the other states. This reduces our precision and increases the likelihood the model will vary by state simply because of fitting to random error. We increase the risk of falsely identifying nonlinearities because we conduct five times as many tests for them. Both approaches have strengths and weaknesses, and it is not possible to say that one is optimal.

The argument that our results might be seriously biased because parents with children who have birthdates near the cutoff may alter their children's school entry date based on their child's abilities makes no sense. First, there are too few violations of the assignment rule to influence the overall results much one way or the other even if these particular cases were tightly clustered around the cut-off. Second, we analyze the data with and without the children whose age violates the assignment rule and don't find any indications of the hypothesized bias. Nor do analyses of the different age bands (from 3 months to 12 months) around the cutoff indicate any such problem.

Interpreting the Results

There are many reasons for variations in the estimated effects across states. Differences in outcomes may be due to differences in the pre-K programs. However, they may also be due to differences in the populations served and differences in the counterfactuals. For example, one

might expect programs to have greater effects on more disadvantaged children. We think that it is particularly important to recognize that the counterfactual is not "no preschool education," and it is possible that the majority of children who did not attend state pre-K attended Head Start or private preschool programs in some states. This makes it difficult to compare our outcomes to those of other programs that had no-preschool education counterfactuals. It also makes it difficult to compare results across the states in our study. The nature of the counterfactual should be kept in mind when judging the importance of the magnitude of our estimated effects. Also, there is no sound empirical basis for relying on Cohen's suggested guide to interpretation of effect sizes. Whether an effect is meaningful or not depends on its value, and not just on its abstract magnitude. Thus, one can measure platinum and brass on the same scale, but the value of an ounce of each remains considerably different. We believe that the effects are large enough to justify the effort, and other analyses have found that our estimated average effect sizes are large enough for these pre-K programs to pass a cost-benefit test.[2] All in all, the conclusion that our study "adds to the evidence" that high quality pre-K can improve learning and development is quite reasonable.

Note: This report is open to public comments, subject to review by the forum moderator. To leave a comment, please send an email to welfareacademy@umd.edu or fill out the comment form at http://www.welfareacademy.org/pubs/early_education/chapter12.html.

[2]Jens Ludwig and Deborah A. Phillips, *The Benefits and Costs of Head Start*, NBER Working Paper Series 12973 (Cambridge, MA: National Bureau of Economic Research, March 2007), http://www.nber.org/papers/w12973.pdf (accessed October 15, 2008).