SCHOOL of PUBLIC POLICY

24
Oklahoma's Universal Pre-Kindergarten

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/chapter24.html.

24
Oklahoma's Universal Pre-Kindergarten

In 1998, Oklahoma established a universal pre-Kindergarten (pre-K) program for all four-year-old children. By 2005-2006, 93 percent of Oklahoma's 543 public school districts were participating in the program and serving about 33,000 children. Oklahoma's program is one of the most "universal" in the country, with about 70 percent of all eligible four-year-olds served, and is noted for its commitment to quality.[1]

William T. Gormley and his colleagues at Georgetown University ("the Georgetown team") conducted two evaluations of the Tulsa Public School District's pre-K program, using a regression-discontinuity design to compare the test scores of children who had attended pre-K with the test scores of children who were about to enter pre-K. The first evaluation took place in 2001, and the second took place in 2003. We focus on the later evaluation as it used the sub-tests of the Woodcock-Johnson Achievement Test which were a significant upgrade over the testing instruments used in the 2001 evaluation and because the later evaluation had a large enough sample to include findings for all racial and ethnic groups.[2] Among four-year-olds who attended the program, the Georgetown team reports that the Tulsa pre-K program produced statistically significant positive effects on children's pre-writing, pre-reading, spelling, math, and problem solving test scores, as measured by the Woodcock Johnson sub-tests; the evaluation did not include any measures of health or behavior.

The Georgetown team conducted a third evaluation, using propensity-score matching and a fixed-effects econometric model to compare the socio-emotional development of children in kindergarten in 2006 who had attended Tulsa pre-K with their classmates who had not. Children who had attended pre-K were less timid, less apathetic, and "less prone to attention-seeking behavior"[3] than their classmates who had not attended Tulsa pre-K.

---

[1]W. Steven Barnett, Jason T. Hustedt, Laura E. Hawkinson, and Kenneth B. Robin, *The State of Preschool 2006* (New Brunswick, NJ: National Institute for Early Education Research, 2006).

[2]William T. Gormley Jr., Ted Gayer, Deborah Phillips, and Brittany Dawson, "The Effects of Universal Pre-K on Cognitive Development," *Developmental Psychology* 41, No. 6 (2005): 881.

[3]Willam T. Gormley, Jr., Deborah Phillips, Katie Newmark, and Kate Perper, Socio-Emotional Effects of Early Childhood Education Programs in Tulsa (paper presented at the Meeting of the Society for Research in Child Development, Denver, CO, April 3, 2009), 23.

However, none of the evaluations reported long-term impacts of the program and it is unclear if the initial positive effects will persist or fade over time. Moreover, it is unclear if the reported findings can be generalized to either the state or national levels. Finally, for the evaluations that made use of regression-discontinuity design, the reported effect sizes are for children with birthdays within twelve months of the cut-off which are somewhat higher than the effect sizes for children with birthdays within three or six months of the cut-off, which calls into question the comparability of the two groups.

Program Design

Program group. Oklahoma's pre-K program targets all four-year-olds across the state. The study participants were drawn from the Tulsa Public School District, which is characterized by a racially and ethnically diverse student population. In the 2003 evaluation, in Kindergarten, 38 percent of children were white, 33 percent were black, 17 percent were Hispanic, 10 percent were Native American, and 1 percent were Asian. In pre-K, 36 percent of children were white, 36 percent were black, 18 percent were Hispanic, 9 percent were Native American, and 1 percent were Asian.[4] The Georgetown team notes that for the kindergarten group, the tested children were more likely to include white children and less likely to include children eligible for free lunch compared to the universe of children in kindergarten. For the prekindergarten group, the tested children were more likely to include black children and children eligible for a free lunch compared to the universe of children in prekindergarten. [5]

For the 2006 evaluation, 35 percent of children in kindergarten were white, 31 percent were black, 23 percent were Hispanic, 10 percent were Native American, and 1 percent were Asian. There was no statistically significant differences in race between children who were assessed and the overall universe of children in kindergarten.

Services. All public school districts in Oklahoma are free to participate in the state's pre-K program, and participating school districts receive state aid for each four-year-old they enroll in a pre-K program. At the time the evaluation was conducted (2002–2003 school year), 91 percent of school districts participated in the program. Oklahoma's pre-K served about 63 percent of all four-year-olds in the state, either directly or through a community-based organization.[6] In 2002–2003, about 56 percent of the programs were half day and 44 percent were full day.

---

[4]Gormley, et al., 2005, 874.

[5]Gormley, et al., 2005, 875.

[6]U.S. Government Accountability Office, *Four Selected States Expanded Access by Relying on Schools and Existing Providers of Early Education and Care to Provide Services* (Washington, DC: GAO, September 2004), 12, http://www.gao.gov/new.items/d04852.pdf (accessed October 15, 2008).

Unlike other early intervention programs, the Oklahoma pre-K program does not provide non-classroom services to either parents or children. Rather, the focus of the program is its educational component. Children spend an average of 3.5 hours in a classroom per day in the part-time programs and 6.5 hours per day in the full-time programs.[7] Teachers are required to have a bachelor's degree and a certificate in early childhood education. Additionally, there is a required ratio of ten students per teacher, and a maximum of twenty students per classroom. Many classrooms have assistant teachers, for whom there are no specific education or training requirements.

The Evaluation. For the 2003 evaluation, the Georgetown team used a regression-discontinuity design to evaluate the Tulsa pre-K program. 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.[8] 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."[9] Estimated program effects are derived through regression analyses using data of participants near the cutoff.[10] Regression-discontinuity designs also need sample sizes that are much larger than random assignment evaluations to achieve the same statistical precision.[11]

The Georgetown team's regression-discontinuity design made use of Tulsa's strict birthday cutoff—children who had birthdays prior to September 1, 2002 were allowed to enroll in pre-K in September 2002 while children with birthdays on or after September 1, 2002 were not allowed to enroll in pre-K until September 2003. The design attempted to control for selection bias by comparing the children near the cutoff in both the program and comparison groups. As the children in both groups were similar in age and demographics, any difference in test scores could be ascribed to the program effect.[12]

---

[7]William T. Gormley, "The Universal Pre-K Bandwagon" *Phi Delta Kappan*, November 2005.

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

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

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

[11]Trochim, 1994.

[12]Gormley, et al., 2005

In September 2003, the Georgetown team administered three sub-tests of the Woodcock-Johnson Achievement Test to 1,567 Tulsa pre-K students and 1,349 kindergarten students who had participated in Tulsa's pre-K program during the previous year. Due to a relatively small number of students that had birthdays near the cutoff, the Georgetown team ran regressions including students with birthdays within three, six, and twelve months of the cutoff, ultimately reporting the program effects from the regression including children with birthdays within twelve months of the cutoff.

In the 2006 evaluation, the Georgetown team used propensity score matching to create the program and control groups as a way to reduce selection bias. In this methodological design, "Members of the treatment and control groups are matched based on having a similar likelihood of being in the treatment group, a measure known as the propensity score, which is estimated from a wide variety of observable characteristics. That is, treated individuals are compared to individuals who 'look' like members of the treatment group, but who did not actually choose the treatment."[13]

In October 2006, kindergarten teachers in the Tulsa school district used the Adjustment Scales for Preschool Intervention (ASPI) to assess the social-emotional development of children in their classrooms. The Georgetown team received completed reports for 77 percent of the kindergarten students. 1,338 had attended Tulsa pre-K and 1,463 had not attended pre-K or Head Start. Children who were born after the September 1 cutoff for entering kindergarten but were still enrolled were dropped from the sample. This included twenty-four children in the program group.

To create the propensity score, the Georgetown team used a logit regression including a number of covariates to estimate the likelihood of children in kindergarten having participated in Tulsa pre-K. Children who were in the program group were then matched to children in the control group who had a similar propensity score. 181 children in the program group did not have "quality matches" and were dropped from the evaluation. Also, in many cases, a child in the control group was matched to more than one child in the program group. Ultimately, 1,133 children were included in the program group and 582 children were included in the control group. The Georgetown team reports that, "children who were black or Native American, whose mother was relatively well-educated, who lived with their biological father, and who had access to the Internet were more likely to be dropped from the treatment group because of poor matching."[14] A subsequent baseline characteristic comparison found that the control group was more likely to be Hispanic and less likely to have internet at home.

Because of the dissimilarity of the two groups, the Georgetown team also used a fixed-

[13]Gormley, Phillips, Newmark, and Perper, 2009, 18.

[14]Gormley, Phillips, Newmark, and Perper, 2009, 21.

effects regression that controlled for a number of demographic factors (including "gender, race, free lunch eligibility, mother's education, whether the child lives with the biological father, and internet access at home")[15] and for the possible differences in how the different kindergarten teachers applied the assessment.

Major Findings

In the 2003 evaluation, among the four-year-old children whose parents chose to place them in pre-K, the Tulsa program had statistically significant effects on multiple areas of their cognitive development, including their pre-reading, reading, pre-writing, and spelling skills, as well as their math reasoning and problem-solving abilities. Children of all races and ethnicities (white, black, Hispanic, and Native American) were shown to benefit, as were children from a variety of economic backgrounds (as measured by their free lunch eligibility).[16]

In the 2006evaluation that focused solely on socio-emotional factors, children that had attended Tulsa pre-K were significantly less likely to be timid, less likely to be apathetic, and more likely to be attentive.

Cognitive. The pre-K program had the largest effect on the Letter-Word Identification test (0.79 SD), followed by the Spelling test (0.64 SD) and the Applied Problems test (0.38 SD).[17] These gains were equivalent to gains of seven months on the Letter-Word Identification test, six months on the Spelling test, and four months on the Applied Problems test, compared to what would have been expected on the basis of aging or maturation alone.[18]

~~The Georgetown team's first evaluation found that the program had statistically significant positive effects for Hispanic and black children, but small to non-existent effects for white children.[19]~~ The 2003 evaluation, ~~which included a larger sample and used a different testing~~

[15]Gormley, Phillips, Newmark, and Perper, 2009, 22.

[16]Gormley, et al., 2005, 872.

[17]Gormley, et al., 2005, 880.

[18]Gormley, et al., 2005, 882.

[19]In the 2001 evaluation, the Georgetown team used a "fixed menu testing instrument" which created an artificial ceiling for scores and limited the ability of the Georgetown team to adequately assess the impact of the program. William T. Gormley, Georgetown University, e-mail message to Douglas Call, March 9, 2008 and William T. Gormley and Deborah Phillips, "The Effects of Universal Pre-K in Oklahoma: Research Highlights and Policy Implications," (CROCUS Working Paper #2, Georgetown University, October 2003)

~~instrument,~~ shows statistically significant effects for children of all races and ethnicities. The largest gains were found for Hispanic children,[20] with effect sizes ranging from 1.5 SD on the Letter-Word Identification test to 0.99 SD on the Applied Problems test. Black children experienced moderate gains on the Letter Word Identification test (0.74 SD) and Spelling test (0.52 SD) and small gains on the Applied Problems test (0.38 SD). While white children also experienced moderate gains on the Letter Word Identification Test (0.76 SD) and the Spelling test (0.72 SD), the gains on the Applied Problems test were not statistically significant.

~~The first evaluation conducted by the Georgetown team found stronger effects for disadvantaged children than for more advantaged ones.[21]~~ In the 2003 evaluation, effects were found across the economic spectrum, as defined by children's free lunch eligibility status. The Georgetown team points to their use of the sub-tests of the Woodcock Johnson Achievement Test in the second evaluation as a possible explanation for this change.[22] Children receiving full price lunch had medium gains on the Letter Word Identification test (0.63 SD) and the Spelling test (0.54 SD), and smaller gains on the Applied Problems test (0.29 SD). Children receiving reduced-price lunch had larger gains on the Letter Word Identification test (1.04 SD) and Spelling test (0.97 SD), but did not have significant gains on the Applied Problems test. Lastly, for those children receiving free lunch, the gains were large to moderate in size for the Letter Word Identification test (0.81 SD) and Spelling test (0.65 SD) and slightly smaller for the Applied Problems test (0.45 SD).

The above effect sizes come from the regressions using the entire data set "ranging from children with birthdays twelve months before the cutoff to children with birthdays twelve months after the cutoff."[23] No effect sizes were calculated for the regressions limiting the data to children with birthdays within six months of the cutoff or within three months of the cutoff, but the Georgetown team does provide co-efficients and standard errors for these regressions. For children with birthdays within six months of the cutoff, all test scores remain statistically significant, but at lower levels of statistical significance. Also, all test score co-efficients are smaller than for children with birthdays within twelve months of the cutoff. For children with birthdays within three months of the cutoff, only the Letter-Word Identification and Applied

---

http://www.crocus.georgetown.edu/reports/effects_of_universal_prek_wp2.pdf (accessed June 27, 2006).

[20]While they describe the impacts across different subgroups, The Georgetown team are careful to note, "one cannot compare the estimated test impacts across subgroups . . . For example, the greater estimated test impacts for Hispanic children relative to Black children does not necessarily imply that a representative Hispanic child will gain more from the program than a representative Black child." Gormley, et al., 2005, 882.

[21]Gormley and Phillips, 2003.

[22]Gormley, et al., 2005, 881.

[23]Gormley, et al., 2005, 878.

Problems sub-test scores are statistically significant; however, the co-efficients for both of these sub-tests are higher than for children with birthdays within six months of the cutoff.

School readiness/performance. Relevant tests apparently not administered or results not reported.

Socioemotional development. In the 2006 evaluation, the Georgetown team found, at the 0.05 level, that children who attended pre-K were significantly less likely to be timid (an effect size of 0.11 SD), less likely to be apathetic (an effect size of 0.12 SD) and more likely to be attentive (an effect size of 0.19 SD) than children who had not attended pre-K. They were also significantly less likely to engage in inappropriate behavior toward the teacher (an effect size of 0.14 SD). At the 0.10 level, children who attended pre-K were also less likely to engage in attention-seeking behavior (an effect size of 0.11 SD) and to engage in inappropriate behavior during a learning task (an effect size of .09 SD).

The Georgetown team also evaluated the socio-emotional effects of Tulsa pre-K on the subgroup of children who qualified for free school lunches. As with the overall sample, at the 0.05 level, children who attended Tulsa pre-K and who also qualified for free school lunches were less likely to be timid (an effect size of 0.19 SD), were more likely to be attentive (an effect size of 0.22 SD), and less likely to engage in inappropriate behavior toward the teacher (an effect size of 0.19 SD). At the 0.10 level, they were less likely to engage in inappropriate behavior during a learning task (an effect size of 0.12 SD).

Health. Relevant tests apparently not administered or results not reported.

Behavior. Relevant tests apparently not administered or results not reported.

Crime/delinquency. Relevant tests apparently not administered or results not reported.

Early/nonmarital births. Relevant tests apparently not administered or results not reported.

Economic outcomes. Relevant tests apparently not administered or results not reported.

Effects on parents. Relevant tests apparently not administered or results not reported.

Benefit-cost findings. The Oklahoma pre-K program cost $3,237 per child for a full-day program (6.5 hours per day) and $1,743 per child for a half-day program (3.5 hours per day).[24] A

[24]William T. Gormley, "The Universal Pre-K Bandwagon" *Phi Delta Kappan* 87 (3) (November 2005): 246–249.

benefit-cost analysis, however, was not conducted.

Overall Assessment

Program theory. The Oklahoma pre-K program is based on the general expectation that early intervention programs promote school readiness and improve developmental outcomes for children. The "universal" aspect of the program aims to provide children with the chance to learn from their peers, as well as from the adult teachers in the classroom. As one Oklahoma pre-K program director notes, "The children of the wealthy arrive right alongside those from low-income and even transient families—with each child teaching the rest new lessons, offering new perspectives and experiences."[25]

Program implementation. The Georgetown team did not discuss the implementation of the Oklahoma pre-K program. They note that at the time the 2003 evaluation was conducted, 91 percent of school districts participated in the program and about 63 percent of all eligible four-year-olds in the state were served.[26] Thus, Oklahoma's program is one of the most "universal" pre-K programs in the nation. It has fairly high quality standards, as judged by the requirement that all Oklahoma pre-K teachers have a college degree and a certificate in early childhood education. Additionally, pre-K teachers are paid at the same rate as other elementary and secondary school teachers in the Tulsa Public School District.

Assessing the randomization. The evaluation did not use random assignment.

Assessing statistical controls in experimental and nonexperimental evaluations. The program group consisted of children whose parents had chosen to place them into the Tulsa pre-K program. Thus, the evaluation estimates the impact of Tulsa pre-K on these children's test scores; it cannot estimate the intent-to-treat effect, as is done in other randomized studies, such as the Head Start Impact Study.

Because the program is voluntary, there is a possibility of selection bias if those children whose parents chose to place them in pre-K differed in some systematic way from those children whose parents chose not to place them in the program. The Georgetown team, in fact, notes in the second evaluation that kindergarten children who attended Tulsa pre-K were less likely to be on a full-price lunch, and more likely to be non-white and on a reduced-price lunch than those kindergarten children who did not attend the program. The pre-K children were also less likely to have a mother who did not finish high school and more likely to have a mother who had

[25]National Institute for Early Education Research, "Power of Universal Pre-K: Oklahoma," in The State of Preschool: 2003 State Preschool Yearbook (New Brunswick, NJ: National Institute of Early Education Research, 2003), 28.

[26]GAO, 2004, 12.

completed some college education. The Georgetown team asserts, "These differences suggest that the cross-sectional analysis might result in biased estimates of the true impact of the Tulsa pre-K program."[27] To control for this potential selection bias, the Georgetown team used a regression-discontinuity design based on Tulsa's strict birthday cutoff qualifications for participation in the pre-K program.[28] They state that "the principal strength of this research design is that both sets of students have parents who affirmatively chose to place them in the TPS pre-K program. This helps to ensure that the students are alike in their talent and motivation—intangibles that are extremely difficult to measure."[29]

However, the Georgetown team does point out that "a potential problem with this strategy is that, whereas the selection criteria may be the same across the 2 years, the control and treatment groups may still have different characteristics."[30] Comparing the two groups of children, they find that most observable characteristics are indeed similar, but that the control group still had a higher percentage of Hispanics and mothers with no high school degree.

The Georgetown team also is able to add covariates to the regressions without significant changes to the results, indicating that their design credibly replicates an experimental design.[31] The covariates included free-lunch eligibility, race/ethnicity, gender, mother's education, and the child's date of birth. The latter covariate allows the Georgetown team to control for the difference in children's scores due to maturation.[32]

Although the Georgetown team runs regressions for children with birthdays within three, six, and twelve months of the cut-off, they only calculate the effect sizes from the regression including children with birthdays within twelve months of the cut-off because there are more data points and are more statistically robust. However, the test score coefficients are smaller, but still statistically significant, for children with birthdays within six months and three months of the cut-off. Though the Georgetown team controlled for children's birthdays, it still seems possible that some of the unexplained variation may be due to maturation, which would explain the smaller

---

[27]Gormley, et al., 2005, 876.

[28]For the 2002–2003 school year, children were only eligible to participate in pre-K if they had been born between September 1, 1997 and September 1, 1998.

[29]William T. Gormley, "Small Miracles in Tulsa: The Effects of Universal Pre-K on Cognitive Development" (paper, National Conference of the Early Childhood Research Collaborative, Minneapolis, MN, December 7, 2007), 4.

[30]Gormley, et al. 2005, p. 876.

[31]Gormley, et al., 2005, 878.

[32]William T. Gormley, Georgetown University, e-mail message to Douglas Call, March 9, 2008.

coefficients for children within three or six months of the cut-off.

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, argues 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.[33] Although the Georgetown team did not use the estimates using children with birthdays within three months of the cutoff, this critique calls into question the validity of using birthdays near the cutoff in a regression-discontinuity design.

In the 2006 evaluation, despite the use of propensity scores and a fixed effects model, there is still the possibility for selection bias. Unlike the second evaluation which compared two groups of children who had their parents choose to enroll them in pre-K, the third evaluation may have problems of selection bias as it may not control for the intangibles of talent and motivation, as mentioned above.

The Georgetown team also expresses concern about the introduction of teachers' bias in the ASPI ratings. They note, "In particular, it is difficult to know how individual biases of various kinds may have affected the ratings. . . .The addition of more objective, observational measures of children's behavior would have strengthened our study still further."[34]

Sample size. The 2003 evaluation used a sample of 1,567 pre-K students and 1,349 kindergarten students from the Tulsa Public School district. The third evaluation used a sample of 1,133 children who had attended Tulsa pre-K and 582 children who had not attended pre-K

Attrition. In the 2003 evaluation, because both groups (treatment and control) were tested only once, attrition is not a factor. In the 2006 evaluation, the Georgetown team originally collected data for 1,338 children who had attended Tulsa pre-K and 1,463 who had not. However, difficulties in finding "quality matches" reduced this number to 1,133 in the program group, or 85 percent of the initial group that was assessed, and 582 in the control group, or 40 percent of the initial group that was assessed.

Data collection. In the 2003 evaluation, the data collection relied on school administrative data and nationally-recognized sub-tests of the Woodcock Johnson Achievement Test. The data sources were appropriate for the questions being studied. In the third assessment, the Georgetown

[33]Michael Puma, Chesapeake Research Associates LLC, e-mail message to Stefanie Schmidt, December 27, 2005.

[34]Gormley, Phillips, Newmark, and Perper, 2009, 32.

team used the Adjustment Scales for Preschool Intervention (ASPI) and school administrative data. The ASPI "consists of 144 statements describing behaviors that children may display"[35] that the teacher uses to assess the student during the assessment period.

Measurement issues. The evaluations rely on standard cognitive, achievement, and socio-emotional tests.

Generalizability. For both evaluations, the samples are racially and ethnically diverse; however, they are only limited to those children who participated in pre-K in the Tulsa Public School District. Thus, there is no way of knowing if the results are generalizable to the remainder of children in the boundaries of the school district that did not participate in pre-K, let alone questions of how representative Tulsa is of the rest of Oklahoma and Oklahoma of the rest of the nation. Also, Oklahoma's quality standards for pre-K make difficult the generalizing of the Tulsa findings to other states with differing quality standards.

In addition, as the children entering Tulsa pre-K scored much lower than the national norm on the three tests, it is unclear if children entering pre-K programs with test scores above the national norm would experience similar positive effects due to pre-K.[36]

Replication. Oklahoma is one of three states (GA, FL) with more that 50 percent of four-year-olds enrolled in pre-K.

Evaluator's description of findings. The Georgetown team is quite positive about both the evaluations' findings. For the 2003 evaluation, they write: "The results provide solid support for the benefits that (school-based universal pre-K) can have on the test scores of young children of differing ethnic and racial groups and from differing socioeconomic backgrounds."[37] They compare the Tulsa pre-K effect sizes to those reported for other state-funded pre-K programs, other pre-K programs generally, high-quality child care programs, and small-scale early intervention programs such as Abecedarian and Perry. They conclude, "The effect sizes reported here fall somewhere in between those of average state-funded pre-K programs and the very best early intervention programs; they substantially exceed those of high-quality child care"[38] and that the lessons learned from their evaluation include that "a well-designed universal pre-K program can produce impressive improvements in student readiness" and that "a well-designed universal pre-K program can benefit children from diverse racial and ethnic backgrounds and from diverse

[35]Gormley, Phillips, Newmark, and Perper, 2009, 14.

[36]Gormley, et al., 2005, 882.

[37]Gormley, et al., 2005, 880.

[38]Gormley, et al., 2005, 881.

social strata."[39] The Georgetown team also demonstrates a high degree of confidence in their regression-discontinuity design and assert that other less rigorous designs, such as naive regressions, actually underestimate the impacts of the pre-K program.[40]

For the 2006 evaluation, they conclude, "The current findings suggest that the children's experiences of positive teacher-student relationships in pre-K may carry over to the kindergarten year as a result of either enhanced social skills or higher expectations of teacher support."[41]

Although the findings may be promising, as mentioned previously, there are still questions about their generalizability.

Evaluator's independence. The evaluators are independent researchers from Georgetown University.

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

Effect sizes. Effect sizes were calculated as the mean differences of the program and control group divided by the standard deviation of the test scores. In the 2003 evaluation, for all children, effect sizes ranged from 0.38–0.79 SD on the cognitive tests. Hispanic children's effect sizes were most pronounced ranging from 0.99–1.5 SD. Also, as mentioned above, effect sizes for children receiving reduced or free school lunches were higher than the average effect size. Under traditional demarcations, the effects for all students range from moderate to large. For the aforementioned subgroups, the effects can be considered large. (See Appendix 1 for a further discussion of effect sizes and their interpretation.)

In the 2006 evaluation, for all children, effect sizes ranged from 0.11–0.19 SD, and for low-income children, effect sizes ranged from 0.19–0.22 SD. Under traditional demarcations, the effect sizes are considered small at best.

However, Thomas Cook, a professor of Sociology at Northwestern University, 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 two programs, the focus of the Tulsa pre-K program on cognitive gains only, and the generalizability of the nationally representative

[39]William T. Gormley, "Small Miracles in Tulsa: The Effects of Universal Pre-K on Cognitive Development" (paper, National Conference of the Early Childhood Research Collaborative, Minneapolis, MN, December 7, 2007), 9.

[40]Gormley, et al., 2005,

[41]Gormley, Phillips, Newmark, and Perper, 2009, 29.

Head Start evaluations compared to the local Tulsa pre-K evaluation.[42]

Sustained effects. The evaluation examined immediate post-intervention impacts for children who had recently completed pre-K. However, the medium- and long-term impacts of the program remain unknown.

Benefit-cost analysis. Apparently not performed.

Cost-effectiveness analysis. Apparently not performed.

[42]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

William T. Gormley, Jr.

One of the most vexing challenges that education policy researchers face is that of selection bias. Unless researchers enjoy the luxury of an experimental research design, they must confront the possibility that their treatment group and their control group will differ in certain unobserved characteristics (e.g., motivation). Even an experimental design is not a panacea.[43] Occasionally, children designated for the treatment do not receive it; more commonly, some children designated for the control group do receive the treatment.

In late 2001, after receiving some early childhood testing data from the Tulsa Public Schools, we decided to apply a regression discontinuity research design to a large-scale pre-K program evaluation. Our strategy was to compare kindergarten students who attended the Tulsa pre-K program the previous year (the treatment group) with students about to begin the Tulsa pre-K program. The big advantage of this research strategy is that children in both groups (or, more precisely, their parents) chose participation in the Tulsa pre-K program rather than various alternatives (Head Start, some other pre-K program, a day care center, a family day care home, or parental care). Thus if more motivated children – or children with certain kinds of parents) are more (or less) likely to enroll in the program, then such children have an equal chance of being included in both the treatment and control groups. This research strategy was possible because Tulsa Public Schools strictly enforced a September 1 cutoff date for pre-K eligibility and because Tulsa Public Schools administered precisely the same test at precisely the same point in time (just prior to the commencement of classes) to incoming pre-K students and to incoming kindergarten students.

Of course, such a research strategy makes sense only if one also controls for the child's precise date of birth. To play it safe, we also chose to control for a wide variety of other demographic variables, though, in principle, such controls should not be necessary if the regression discontinuity design works as intended. That is because children in the treatment and control groups should have similar demographic characteristics, except for age.

Our initial application of the regression discontinuity design to the Tulsa data[44] worked

[43]Jeff Grogger and Lynn Karoly, *Welfare Reform: Effects of a Decade of Change* (Cambridge, MA: Harvard University Press, 2005).

[44]William Gormley, Jr., and Ted Gayer, "Promoting School Readiness in Oklahoma," *Journal of Human Resources* 40 (Summer 2005): 533-58.

extremely well, from a methodological perspective, except that the test data suffered from "ceiling effects" problems because the data we received from the school district came from a 26-item test. For middle-class students, in particular, this posed a problem because students might have scored even higher had there been more items in the test. In short, some test scores were prematurely truncated.

To solve that problem, we persuaded the Tulsa Public Schools that their teachers should administer three subtests of the Woodcock Johnson Achievement Test in August 2003. These subtests, suitable for four-year-olds, five-year-olds, and older children, included an abundance of test questions so that ceiling effects would not be a problem. The subtests were also nationally normed, to facilitate comparisons with other students. Ultimately, we were able to compare 1,567 pre-K children and 1,461 kindergarten children who had just attended pre-K the previous year. We found test impacts of .79 of a standard deviation for the Letter-Word Identification Test, .64 of a standard deviation for the Spelling Test, and .38 of a standard deviation for the Applied Problems Test.[45]

When applying the regression discontinuity framework to test score data or other data, researchers must decide whether to include all data points available or focus on a narrow slice of data points around the cutoff point or something in between. There are trade-offs here, on which researchers can reasonably disagree. Estimates based on a narrower slice of data are typically less biased but also less efficient.

In our various publications, we have offered a range of options with respect to children's age. Specifically, we have analyzed data for all children with birthdays that do not exceed the cutoff birthday in either direction by 12 months, then 6 months, then 3 months. We have invited readers to choose based on their own attitudes toward the bias-efficiency trade-off. Our own preference is to use all the data points (12 months), for the sake of efficiency. While the coefficients vary somewhat as one shifts from 12 months to 6 months to 3 months, there is not a steady upward or downward trend. Whatever one's preference on this matter, the bottom line is that the Tulsa Public Schools pre-K program substantially improves children's cognitive development.

It is always a good idea to raise questions about a new methodology or the application of a familiar methodology to a new setting. In addition to their own comments, Besharov et al. cite a concern attributed to Michael Puma: the possibility that parents of "more able" children who have birthdays near the cutoff will be more likely to enroll their child in pre-K earlier, while parents of "less able" children who have birthdays near the cutoff will be more likely to enroll their child in pre-K later. Because the pre-K program is available only to four-year-olds, a more realistic

[45]William Gormley, Jr., Ted Gayer, Deborah Phillips, and Brittany Dawson, "The Effects of Universal Pre-K on Cognitive Development," *Developmental Psychology* 41 (November 2005): 872-84.

concern would be if parents of "more able" children who have birthdays near the cutoff were more likely to enroll their child, while parents of "less able" children who have birthdays near the cutoff were less likely to enroll their child at all.

If this scenario were correct, then it might pose a problem for the regression discontinuity design because relatively young pre-K alumni in the treatment group would be "more able" than relatively old pre-K entrants in the control group with whom they are being compared. If Puma's concern were warranted, we might expect to see more children with well-educated mothers or more middle-class children just to the right of the September 1 cutoff point (i.e., very young pre-K alumni) than just to the left of the September 1 cutoff point (i.e., very old pre-K entrants). However, there are no differences in the likelihood of being poor or middle-class at the cutoff.[46] Moreover, very young pre-K alumni are actually *more* likely to have a mother who did not graduate from high school than very old pre-K entrants (no statistically significant differences in our other mother's education variables). If anything, this would seem to make it slightly *more* difficult to discern positive effects at the cutoff point, using the regression discontinuity design.

Of course, Tulsa is only one community, and the Tulsa Public Schools pre-K program is only one pre-K program. We chose Oklahoma because it has the highest penetration rate of any pre-K program in the country, and we chose Tulsa because it is the largest school district in Oklahoma. We have always pointed out that the Tulsa Public Schools pre-K program is a high quality program, with a college-educated, early-childhood-certified teacher in every classroom. We do not assume that every pre-K program is as good as Tulsa's. In fact, we have documented differences between Tulsa Public Schools pre-K classrooms and other school-based pre-K classrooms in 11 states.[47]

Nevertheless, it is important to stress that significant improvements in children's cognitive development have been discovered in other evaluations of state pre-K programs. Since our initial regression discontinuity paper appeared in 2003, other scholars have chosen to use the same analytical strategy to assess the effects of state pre-K programs in five states, including Oklahoma.[48] In Oklahoma as a whole, pre-K participation has produced impressive results. The

---

[46]William Gormley, Jr., Ted Gayer, Deborah Phillips, and Brittany Dawson, "The Effects of Universal Pre-K on Cognitive Development," *Developmental Psychology* 41 (November 2005): 877.

[47]Deborah Phillips, William Gormley, Jr., and Amy Lowenstein, "Classroom Quality and Time Allocation in Tulsa's Early Childhood Programs" (paper, Biennial Meetings of the Society for Research in Child Development, Boston, MA, March 30, 2007).

[48]W. Steven Barnett, Cynthia Lamy, and Kwanghee Jung, *The Effects of State Pre-Kindergarten Programs on Young Children's School Readiness in Five States* (New Brunswick, NJ: National Institute for Early Education Research, December 2005), http://nieer.org/resources/research/multistate/fullreport.pdf (accessed July 15, 2008); and Vivian Wong, Thomas Cook, W. Steven Barnett, and Kwanghee Jung,"An Effectiveness-Based Evaluation of Five State Pre-Kindergarten Programs," *Journal of Policy Analysis and Management* 27 (Winter

same is true of the other four states: Michigan, New Jersey, South Carolina, and West Virginia.

The moral of this story is not that any old pre-K program will produce miraculous results but rather that Tulsa is not unique. Other high-quality state-funded pre-K programs in other jurisdictions have also enhanced the school readiness of large numbers of children.

2008): 122-54.

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/chapter24.html.