

State-Funded Preschool:

Analyses of Third-Grade ITBS Results of Children Enrolled in Iowa's 2007-08 Statewide Voluntary Preschool Programs



Arie van der Ploeg
American Institutes for Research

Valeriy Lazarev
Empirical Education Inc.

This publication was prepared by the American Institutes for Research, 20 N. Wacker Drive, Suite 1231, Chicago, IL 60606. This publication is in the public domain. Authorization to reproduce in whole or in part for educational purposes is granted.

	Page
Introduction	4
Background.	6
Issues Constraining Analytic Choices	7
Design Options	8
Visualizing the Data	9
Analyses With Matched Comparison Groups	11
Summary	15
References	16

Introduction

Every high-performing education system in the world has a quality early childhood education component. In Iowa, the Statewide Voluntary Preschool Program for 4-year-old children was adopted by state legislators in 2007 as an important part of the state's comprehensive early childhood effort. The goal was to provide another opportunity for young children to access quality preschool and to enter school ready to learn.

This report was commissioned by the Iowa Department of Education to examine the longitudinal impact of the preschool program. The study was completed by the American Institutes of Research (AIR), a nationally recognized organization with expertise in education and social science research. The Department sought out AIR as an independent and objective resource to evaluate the outcomes of the preschool program at no charge.

This report examines the impact of the preschool program statewide by tracking the progress and achievement of its young participants over time. Specifically, the evaluation design examines the third-grade assessment results of the first cohort of children who participated in the preschool program during the 2007-08 school year. The goal is to determine whether the preschool program had a lasting impact on the subsequent test scores of this first cohort of students.

It must be understood that this report has both strengths and limitations. An important strength is that the report creates a matched control group, which provides a macro-level analysis to determine whether or not the preschool program had an impact on student achievement results later on. At the same time, the report has a limited focus, using student achievement results as the sole predictor of lasting educational impact. It also should be noted that school districts had a short window of time to establish quality preschool programs during the first year of the preschool program, which launched just a few months after Iowa legislators created it in May 2007.

Earlier reports released by the Department show initial and modest gains in achievement results for the preschool program on a kindergarten readiness assessment. This report concludes that this impact appears to dissipate by third grade in both reading and mathematics. These findings are consistent with the recent evaluation of the federal Head Start program (Puma, et.al, 2012).

While this report's findings show no impact on students' test results four years after preschool, the report does not suggest the preschool program had no impact on participants. Multiple studies have shown positive outcomes for students that reinforce the need for early childhood education programs. Examples of these outcomes include non-cognitive factors such as socialization, social mobility, and persistence (Heckman, 2007; Barnett & Belfield, 2006).

This report should be considered initial feedback. The Department recognizes the need for deeper analysis that factors in other outcomes, such as non-cognitive impact, and also differentiates between district-level programs, which may vary in quality, standards, and length of time.

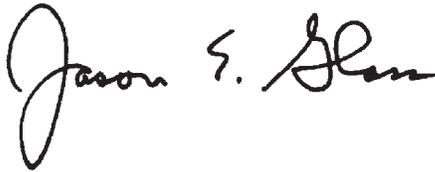
We must give education leaders, policymakers, and parents the tools to take a critical look at early



childhood education programs in place in Iowa school districts and to find answers to the following questions:

- Is the length of the program sufficient to provide results that will translate into kindergarten readiness?
- Are the program standards clearly articulated and aligned?
- Are supports and professional development opportunities available for teachers in these programs?
- Have particular student populations, such as students from economically disadvantaged backgrounds, been targeted for participation?

The goal is not to repeal this preschool program. Instead, the Department aims to make the improvements that will ensure this program is effective in meeting the needs of Iowa students and in raising student achievement.

A handwritten signature in black ink that reads "Jason E. Glass". The signature is written in a cursive style with a large initial 'J' and 'G'.

Jason E. Glass, Ed.D.
Director, Iowa Department of Education

Background

In August 2012, the American Institutes for Research (AIR) received an inquiry from Jay Pennington, Chief of the Bureau of Information and Analysis Services within the Iowa Department of Education (IDE). He sought AIR's objective expertise to recommend and implement a rigorous design under a quick turnaround timeline to analyze the academic performance in third grade of students who four years earlier had participated in a voluntary state-funded program for four-year-olds. A memorandum of understanding and a data sharing agreement were executed on October 11, 2012.

AIR received the first data files several days later. In all, IDE provided the following data sets:

- A file identifying the districts that were awarded Statewide Voluntary Preschool Program (SWVPP) grants for 2007–08
- A file of administrative data for students enrolled in prekindergarten programs statewide in 2007–08
- A file of administrative data for students enrolled in kindergarten statewide in 2008–09
- A file of administrative data for students enrolled in third grade statewide in 2011–12
- A file of ITBS results for third-grade students statewide in 2011–12
- Three files of district ITBS means, by test period, school year, and student subgroups, 2008–12

Over the past decade, some of Iowa's schools and school districts have conducted a variety of prekindergarten instructional programs, some supported by state funds, some by local funds, and others privately funded. Not all districts supported such programs. Programs varied with respect to the services provided as well as their length and frequency. In 2007, the Iowa Legislature authorized funds for a new SWVPP for the 2007–08 academic year. Districts were asked to prepare proposals outlining their plans. Funds were awarded to 66 districts (mergers reduced the number to 62 by 2011–12).

The research literature on early childhood interventions and their impact on academic performance is easy to summarize (Heckman & Masterov, 2007; Jencks, 2013; Reynolds, Temple, Ou, Arteaga, & White, 2011). Compared with nonparticipants, participants in early childhood typically experience modestly positive impacts on academic outcomes by year's end although that advantage fades within several years. However, strong evidence now exists that early childhood and prekindergarten programs have visible distal effects, such as reducing grade repetition, disability diagnoses, and health problems, while increasing high school graduation, college attendance, and employment. Evidence also is beginning to accumulate that more exposure to early childhood programs (more days in the year, more hours to the day) increases impact (Loeb, Bridges, Bassok, Fuller, & Rumberger, 2007).

Issues Constraining Analytic Choices

The original district SWVPP proposals were not available to this project. Therefore, it was not possible to capitalize on between-district variations with respect to program design elements such as whether and how SWVPP students were recruited, whether and what academic content was taught, or length of instructional day. The 2007–08 administrative file identified the children enrolled in SWVPP and other prekindergarten programs. The 2008–09 kindergarten file included fields that recorded parents' responses about their child's preschool experiences. The presence of these data elements offered the possibility of comparisons among students enrolled in SWVPP, other public prekindergarten programs, other prekindergarten programs (private or out of state), and students without prekindergarten experience.

IDE asked that academic performance be operationalized using results obtained from the state's testing program. The state's annual academic and accountability testing is conducted by the Iowa Testing Programs research center at the University of Iowa, working directly with districts and schools. The program allows districts to administer the ITBS at the time of their preference. IDE sorts scores into three periods (fall, midyear, and spring) and implements statistical adjustments to produce comparable standard scores and proficiency ratings each year. The analyses described later use only the comparable standard scores. Proficiency ratings collapse much information and present relatively intractable computational problems, given their nonequal interval property.¹ Adding complexity is the fact that the 2011–12 ITBS administration introduced a revised form of the tests with new norms. Scores from 2008–11 are said by Iowa Testing Programs to be comparable year-to-year, but this is not so for the 2012 scores. In addition, results obtained from the few students who took the Iowa Alternate Assessment in 2011–12 were excluded from the analyses.

¹Recent work is identifying methodologies that reproduce some full-distribution estimates quite well (Ho & Reardon, 2012). However, scale scores remain preferable.

Design Options

A stronger, less vulnerable design would more clearly state whether some implemented policy (SWVPP) produced a distinct outcome. A conceptual model to accomplish this goal requires a comparison of the distribution of the values of an outcome for districts, schools, and students participating in SWVPP to the distribution of outcomes for the same districts, schools, and students with SWVPP not implemented. Superficially, this appears to be an impossible task: something cannot be and not be simultaneously. All available strategies to generate solutions for this quandary target the identification of suitable counterfactuals (Murnane & Willett, 2011).

To assess causal impact, randomized experiments are considered optimal (Shadish, Cook, & Campbell, 2002). Regression discontinuity designs are not experiments per se but do produce impact estimates similar to those from experiments, at least among units located near the cutpoint on selection measures (Gleason, Resch, & Berk, 2012). Among quasi-experimental designs, interrupted time series, given optimal conditions, are frequently able to reproduce experimental results closely (Bloom, 2003). However, the SWVPP intervention occurred four years before analysis was requested, and randomization now is not an option. The lack of a criterion against which districts were selected to the SWVPP program or students were selected to participate removes a regression discontinuity design from consideration. No time-series data are available for the children who participated in SWVPP in 2007–08, either for the years before participation or for the years between SWVPP participation and Grade 3, so an interrupted time-series approach is not feasible.

For all the remaining quasi-experimental designs, defining and constructing appropriate comparison groups is a critical concern; a goal is to approximate the comparison a randomized experiment would have made (Rubin, 2008). Comparing Grade 3 ITBS results of former SWVPP students to results of all other Iowa third graders is not a sound comparison. Four years earlier, some districts sought SWVPP under the Iowa's competitive grant process; others did not. That simple difference may be sufficient to bias subsequent comparisons.

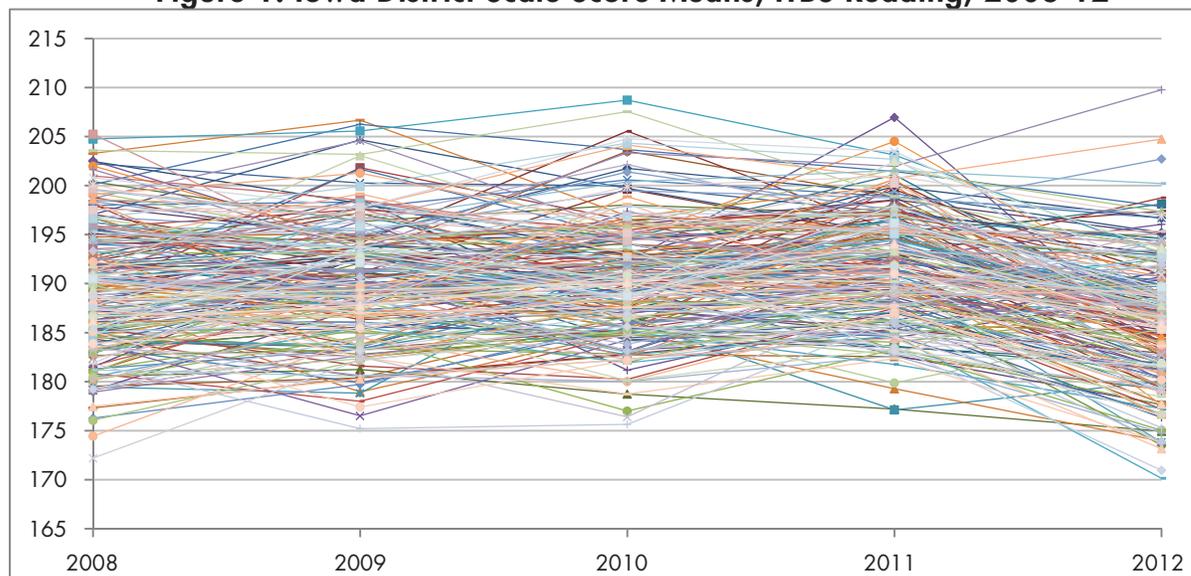
A next-best option is to employ a matching to construct comparison groups that are “balanced,” that is, are similar on many characteristics except SWVPP participation. But what characteristics should be part of this balancing? Students are many-faceted, as are their classrooms and schools. Unfortunately, there is yet no consensus how matching should be done or how to measure its success (Sekhon, 2011). Given the data available and the questions of interest, we adopted a multivariate nearest-neighbor matching procedure with regression adjustments.² To assure a sufficiency of data to the procedure, we opted to match SWVPP students to nonparticipating students rather than match SWVPP districts to nonparticipating districts. After these procedures, the matched students are considered essentially equivalent given the data elements used in the calculations, differing only in preschool program participation. It will be these more sharply defined differences that lend potency to the results obtained.

²Matching was exact on student characteristics and program features. Any one SWVPP student may have no or one or more than one matched non-SWVPP counterparts (see, e.g., Gu & Rosenbaum, 1993). Calculating standard errors under such conditions is complex. We rely instead on multiple point estimates.

Visualizing the Data

The first ITBS results from 2007–08 SWVPP-participating students occurred when they became third graders in the 2011–12 school year. Whether districts chose to enroll all their 4-year-old students or only some, and, if only some, whether they were the neediest or the first to enroll, is not known. Regardless, it is not unreasonable to begin the data exploration by visualizing the pattern of district third-grade ITBS means over time. If SWVPP had strong and lasting effects on students' academic performance, a visible bump-up of district means in 2011–12 would suggest some confirmatory evidence. Figure 1 is such a plot.

Figure 1. Iowa District Scale Score Means, ITBS Reading, 2008-12



The trend lines of Figure 1 permit several conclusions.³ District reading scale score means vary from year to year, with gains typically following losses and vice versa. This variation is limited to a narrow band for most districts, in the range of approximately 190, plus or minus just seven or eight scale score points. The overall mean is less in 2012, and the range of district means somewhat wider—as might be expected for a new test. When viewing the years 2008–11 only, a mild upward tendency appears visible.

But, did SWVPP districts perform better than non-SWVPP districts? To explore that question using the Figure 1 summary data, lines of best fit may be drawn through each district trend line, for the period 2008–11 and the period 2008–12 for reading and for mathematics separately. Table 1 displays the results. It averages the slopes of the trend lines and compares these for the two types of districts in two periods, for ITBS reading and mathematics.

³The chart displays about 80 percent of all Iowa school districts in 2011–12. To enhance legibility, the smallest districts were excluded. Figure 1 includes all districts with at least 24 tested students in Grade 3. A plot for mathematics produces a visually nearly indistinguishable graph.

Table 1. District Time-Series Slopes Compared

	SWVPP		Non-SWVPP	
	Reading	Mathematics	Reading	Mathematics
2008–12 slope	-1.07	0.67	-0.69	0.52
2008–11 slope	0.43	0.57	0.25	0.59
Difference between slopes	-1.49	0.10	-0.94	-0.07
Difference between SWVPP and non-SWVPP	-0.55	0.17		

Note: SWVPP = Statewide Voluntary Preschool Program

For the 2008–11, the slopes are mildly positive in both subjects (scores are improving a little over time) and are somewhat larger for SWVPP districts in reading and about the same in mathematics. For 2008–12, the effect of the new norms is large enough in reading to turn the overall trend negative but not so for mathematics. Some quick subtractions produce a quick back-of-the-envelope guesstimate of the size of the change in scores over time, net of the impact of the 2012 test changes. That difference appears in the bottom row of the table. It suggests that SWVPP districts were losing about one-half scale score point per year in reading compared with the progress of non-SWVPP districts; in mathematics it suggests no difference between SWVPP districts and non-SWVPP districts. These figures are small. Taking into account that they are produced by a rough calculation, it is reasonable to conclude that Figure 1 and Table 1 present, at best, inconclusive evidence about the impact of SWVPP on future third-grade academic performance.

Analyses with Matched Comparison Groups

Reliance on back-of-envelope calculations is not recommended. To design a more focused and defensible analysis, we switch perspectives and turn to the individual student data files. The 2007–08 prekindergarten data file identified records for 4,828 students enrolled in 66 districts receiving SWVPP services and 3,772 students from 230 districts who were not receiving SWVPP services. In addition, the 2008–09 prekindergarten data file identified 15,647 students whose records indicated that their parents said they had participated in a preschool program the prior year.⁴ The 2011–12 file contained records for 35,179 third-grade students. This number reduced to 30,449 records with matches to the 2008–09 kindergarten, the 2007–08 prekindergarten file, or both. Of these records, 1,228 did not have ITBS reading and mathematics third-grade scores. Another 4,112 student records were set aside because they were missing data for the data elements used in the matching procedure, leaving an analytic data set of 25,115 records.

The students in this data set were classified into the categories enumerated in Table 2. There were about a thousand students among the SWVPP and other public prekindergarten participants who were enrolled for less than three weeks. These students were deleted from analyses. The “Other Iowa public prekindergarten” group represents a potentially competitive comparison for SWVPP, if the question is, did one or another program perform better. However, even less is known of the details of the services in this category than of the details of the conduct of SWVPP. The “Other” category represents kindergarten students whose parents reported their children had been enrolled in a preschool program but were not so identified in Iowa administrative records. The data do not make clear if that experience was private, secular, or religious, in Iowa or elsewhere. It is therefore not clear how to interpret comparisons including this group. The “No Prekindergarten” category includes kindergarten students known not to have prior academic experiences. This category supports comparisons to address questions as to whether SWVPP specifically adds value to the school experience of Iowa students.

Table 2. Student Groupings to Be Used in the Analysis⁵

Student Prekindergarten Status	Attendance	Potential Sample	Attendees With ITBS Grade 3 Scores and Data for Matching	
			Sample N	Matched N
SWVPP	< 15 days	550	0	0
	> 15 days	4,278	3,069	2,802
Other Iowa public prekindergarten	< 15 days	443	0	0
	> 15 days	3,329	2,858	1,013
Other*		15,467	14,067	2,038

⁴There were some discrepancies between this 2008–09 information and what was recorded in the 2007–08 data file. We resolved such discrepancies by giving priority to the 2007–08 file. The data also identified 6,202 kindergarteners with no record of prior school experience.

⁵Des Moines is sui generis in Iowa, its only large city. There are no other urban environments within Iowa that match it well. Therefore, some analyses do not include Des Moines students. Although this sharply reduces the number of SWVPP students, by more than 500, it improves the quality of the comparisons in the remaining sample.

Student Prekindergarten Status	Attendance	Potential Sample	Attendees With ITBS Grade 3 Scores and Data for Matching	
			Sample N	Matched N
No prekindergarten		6,202	5,121	4,514
Total		30,449	25,115	10,367

Note: ITBS = Iowa Tests of Basic Skills; SWVPP = Statewide Voluntary Preschool Program.

* Denotes children not identified in Iowa administrative records as having attended preschool but whose parents' reported that their children had experienced preschool. Presumably, this includes a broad variety of non-public or out-of-state experiences.

Although details about each district's SWVPP implementation were not available, the SWVPP attendance data made clear that some districts' programs operated for one semester, and other districts' programs operated for two semesters. This length difference may be consequential. Therefore, program length, part year or full year, was one of the variables used to identify the students to be matched to each of the four groups defined in Table 2. The other matching variables were all student characteristics: gender, ethnicity, special education status, free or reduced-price lunch status, and school transfers.⁶ The matching procedure was executed twice, once seeking matches anywhere in the state and once seeking matches within the SWVPP student's own district only.

Summaries of the results appear in Tables 3 and 4. The cells express the difference between the SWPP students and a matched comparison group. In each case, the table contains the difference between the mean scale scores of the two groups, followed by the p value of that difference, the scale score difference expressed in an effect size metric, and lastly expressed as a percentile difference. The effect size used is the simple standardized mean difference, the difference in scale score means divided by the pooled within-group standard deviation. A value of 0.5 expresses a difference equal to half a standard deviation. That would normally be considered a large difference. In educational research, an effect size of approximately 0.2 is often taken as the dividing line between an inconsequential and a consequential difference. This is not a hard-and-fast rule, however. The percentile difference attempts to capture the expected change that would occur to an average comparison group member if he had participated in the intervention, that is, SWVPP (see, e.g., What Works Clearinghouse, 2011).

Table 3 reports these statistics to compare the Grade 3 ITBS results obtained by former SWVPP participants and those from matched students in the same districts who had not participated in any prekindergarten program. Although the scale score difference carries a negative sign, that SWVPP students' mean Grade 3 ITBS score was lower than that of the matched comparison group whose students had no similar experience, the result was not close to being statistically (or practically) significant, with effect size and percentile difference metrics showing no difference at all between the two groups.

⁶This procedure guaranteed exact equality between the matched groups on these characteristics. Nevertheless, full equality cannot be guaranteed. For instance, the matched SWVPP students started in districts where 2007–08 ITBS reading and mathematics proficiency rates were approximately five scale score points lower than in the matched non-SWVPP students' districts.

Table 3. Grade 3 ITBS Results for Students Enrolled in SWVPP Compared With Students With No Prekindergarten Experience

Metric	Mathematics	Reading
Scale score point difference	-0.08	-0.14
p <	0.85	0.78
Effect size	0.00	0.01
Percentile difference	0	0

Table 4 reports these statistics for the comparison of SWVPP participants to matched students statewide who had no prekindergarten experience. Two sets of results appear: with and without Des Moines. Given the fact that it is difficult to compare Des Moines to any other city in Iowa, the authors highlight the “Without Des Moines” comparison. Students who had no prekindergarten experience outperformed SWVPP students as third graders by about four scale score points on the ITBS in both reading and mathematics. Both differences reach standard levels of significance, and effect sizes are in the range of one fifth of a standard deviation. The numbers suggest that an Iowa 4-year-old would have done better as a third grader not to have enrolled in SWVPP. That avoidance would have shifted the average SWVPP participant upward six to nine percentile points on the third-grade test. The differences are less pronounced if the Des Moines SWVPP program students are included in the comparison.

Table 4. Grade 3 ITBS Results Compared for Students Enrolled in SWVPP and Other Public Prekindergarten Programs

Metric	With Des Moines		Without Des Moines	
	Reading	Mathematics	Reading	Mathematics
Scale score point difference	-1.45	-1.57	-4.15	-3.55
p <	0.16	0.25	0.00	0.01
Effect size	0.08	0.07	-0.24	-0.16
Percentile difference	-3	-3	-9	-6

But, the matched students in this comparison are drawn statewide, not necessarily from within the SWVPP districts. An alternative analytic approach would take advantage of the nested structure of students in districts by adopting a hierarchical linear modeling (HLM) strategy (Raudenbush & Bryk, 2002).⁷ Under this strategy, the mathematics results, displayed in Table 5, show no meaningful differences, and the reading results are marginal, favoring the other public prekindergarten programs and not SWVPP.

⁷The HLM approach was not used in the prior analyses because matching was not within district. Even in within-district matching case, the two-level model used (students within districts) omits several levels; in reality, SWVPP students were nested within teacher with SWVPP program within school within district.

Table 5. Grade 3 ITBS Results Compared for Students Enrolled in SWVPP and Other Public Prekindergarten Programs, Under an HLM Analytic Strategy

Metric	Mathematics	Reading
Scale score point difference	-0.11	-1.09
p <	0.45	0.1
Effect size	0.01	0.05
Percentile difference	0	2

Finally, Table 6 reproduces the results for the comparison to the “Other” category—kindergarten students whose parents say they received preschool experiences but of which Iowa has no record. This too shows SWVPP students performing less well, the comparisons reach significance, and the effect size and percentile metrics suggest they are of at least marginal consequence.

Table 6. Grade 3 ITBS Results Comparing Students Enrolled in SWVPP to Students Whose Parents Say They Experienced Some Preschool Program Other Than an Iowa Public School Program

Metric	Mathematics	Reading
Scale score point difference	-2.15	-4.11
p <	0.00	0.00
Effect size	-0.12	-0.18
Percentile difference	5	7

Summary

What have we learned? The literature tells us that early childhood academic experiences produce short-term academic growth, visible on commonly used tests and assessments, but that these soon fade. That pattern seems to be replicated for SWVPP in Iowa. Conversely, the same literature on early academic experiences tells us that there are numerous nonacademic effects that show much later, in young adulthood and beyond, and that these effects are pronounced, the more so for children from disadvantaged families.

What more have we learned? Carefully designed studies that focus close attention on meaningful comparisons executed with rigorous detail almost always provide more enlightenment than less rigorous approaches. If a question matters, then methods matter.

In sum, what may be said about the performance of former SWVPP students when they reach third grade? The evidence presented here can be interpreted to support the claim that SWVPP students when they reach third grade perform much like similar students without preschool programs. Nevertheless, most of the comparisons carry negative signs, implying SWVPP students do less well than matched students in third grade. But, the differences are consistently small and statistically insignificant. And, it must be kept in mind that in 2007–08, the SWVPP districts were more diverse and somewhat lower performing than other Iowa districts.

It is unrealistic to expect a study such as this to produce unequivocal evidence of significant positive impact. There are two major reasons for this, one having to do with implementation and one with data limitations. Strong positive effects are unlikely given only nine months of implementation: Educational programs are complex, require substantial human capital and human resources, and are typically implemented unevenly. Not all the SWVPP programs opened on the first day of the school year. Second, the historical data available support only limited opportunity for rigorous analytic design. The choices available to families will have varied among districts and, even where SWVPP was the only option, family self-selection to the program will have played a major role in decisions to enroll a child. In districts where other preschool options existed, wealthier families were presumably more likely to enroll children in established private preschools. Such considerations suggest that the observed higher third-grade outcomes may reflect factors other than the quality of SWVPP. Although we implemented several strategies to match SWVPP students to appropriate comparison students, the accuracy of these matches is limited by the small set of student and district characteristics used.

The fact that SWVPP programs continued after 2007–08 opens an avenue to improved analytic designs. Over the next several years, it will be possible to leverage successive cohorts of SWVPP students, including student cohort(s) from the years preceding the introduction of SWVPP and after its first year. Incorporating data on the preprogram cohorts allows comparisons with within-districts matches, comparing students entering SWVPP to those who might have but could not because the program did not exist for them. Data from later years would allow estimation of the effect of more mature program implementations in subsequent years. Finally, using more years of outcomes—for example, fourth-grade scores in addition to third-grade scores—would increase accuracy and precision by reducing the impact of measurement error inherent in a single measure of student achievement.

References

- Bloom, H. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reforms, with applications to a study of Accelerated Schools. *Evaluation Review*, 27, 3–49.
- Gleason, P., Resch, A., & Berk, J. (2012). *Replicating experimental impact estimates using a regression discontinuity approach* (NCEE Reference Report 2012-4025). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.
- Gu, X., & Rosenbaum, P. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4), 405–420.
- Heckman, J., & Masterov, D. (2007). *The productivity argument for investing in young children* (NBGER Working Paper 13016). Cambridge, MA: National Bureau of Economic Research.
- Ho, A., & Reardon, S. (2012). Estimating achievement gaps from test scores reported in ordinal "proficiency" categories. *Journal of Educational and Behavioral Statistics*, 37(4), 489–517.
- Jencks, C. (2013, March 7). *Learning from puzzles: Test scores, habits, behavior, and schooling*. Keynote address to Society for Research on Educational Effectiveness, Washington, D.C.
- Loeb, S., Bridges, M., Bassok, D., Fuller, B., & Rumberger, R. W. (2007). How much is too much? The influence of preschool centers on children's social and cognitive development. *Economics of Education Review*, 26, 52–66.
- Murnane, R., & Willett, J. (2011). *Methods matter: Improving causal inference in educational and social science research*. New York: Oxford University Press.
- Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Reynolds, A., Temple, J., Ou, S-R, Arteaga, I., & White, B. (2011). School-based early childhood education and age-28 well-being: Effects by timing, dosage, and subgroups. *Science*, 333(6040), 360–364.
- Rubin, D. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, 2(3), 808–840.
- Sekhon, J. (2011). Multivariate and propensity score matching software with automated balance optimization: The Matching package for R. *Journal of Statistical Software*, 42(7), 1–52.
- Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- What Works Clearinghouse. (2011). *Procedures and standards handbook (version 2.1)*. Washington, DC: Author.