

High-Stakes Choice: Achievement and Accountability in the Nation's Oldest Urban Voucher Program

John F. Witte

University of Wisconsin–Madison

Patrick J. Wolf

University of Arkansas

Joshua M. Cowen

Michigan State University

Deven E. Carlson

University of Oklahoma

David J. Fleming

Furman University

This article considers the impact of a high-stakes testing and reporting requirement on students using publicly funded vouchers to attend private schools. We describe how such a policy was implemented during the course of a previously authorized multi-year evaluation of the Milwaukee Parental Choice Program, which provided us with data on voucher students before and after the reform, as well as on public school students who received no new policy treatment. Our results indicate substantial growth for voucher students in the first high-stakes testing year, particularly in mathematics, and for students with higher levels of earlier academic achievement. We discuss these results in the context of both the school choice and accountability literatures.

Keywords: *school vouchers, high-stakes testing, accountability*

Introduction

THE STATE of Wisconsin operates the nation's oldest school voucher system for low-income students.¹ Beginning as a small pilot program in 1990, the Milwaukee Parental Choice Program (MPCP) grew to serve 20,996 students in 107 of the city's private schools by 2010—enrolling nearly one fifth of the city's K–12 students. Since its inception, the MPCP has been funded by state and local taxpayers and administered by the Wisconsin Department of Public Instruction (DPI), the state's central educational agency.

For the first two decades of the MPCP, participating private schools were not subject to any significant testing or reporting requirements.

This changed, however, with the passage of legislation that—beginning in the 2010–2011 school year—required private schools participating in the MPCP to annually test all students receiving vouchers in Grades 3 to 8 and 10 with the reading and math portions of the Wisconsin Knowledge and Concepts Examination (WKCE), the same test used by public schools in the state to meet federal and state accountability requirements. The legislation further required all test results to be submitted to the DPI for public reporting. This policy change occurred toward the end of our 5-year, state-mandated evaluation of the MPCP, providing us with pre-reform test scores as well as scores from one post-reform follow-up year.

With these data, we estimate the impact of the private school high-stakes testing policy on student achievement outcomes using a rigorous quasi-experimental approach—the difference-in-differences estimator. In doing so, we provide among the first evidence on the effect of applying a high-stakes testing policy to private schools serving students receiving publicly funded vouchers.

Our general finding is that implementation of the new high-stakes testing regime had a significant positive impact on the achievement scores of voucher students in the first year that the law was in effect. We also provide some evidence that these effects vary across racial/ethnic and prior achievement sub-groups, but in nearly all cases, they remain positive and statistically significant at least in mathematics. Taken as a whole, the results suggest that—even without attaching explicit sanctions for poor performance—applying testing and public reporting requirements to private schools will improve test scores in that sector either through enhanced test preparation or through meaningful gains in educational quality itself.

We proceed by first providing brief background on both private school vouchers and high-stakes testing policies—two literatures that, to this point, have had little overlap. We then transition to a description of the MPCP context and provide an overview of the evaluation that generated the data we draw upon and detailing the testing policy we analyze. Next, we describe the specific data set and sample that serves as the basis of our empirical strategy, which employs difference-in-differences techniques. Finally, we present the results of the analysis and close the article with a discussion of their implications for research and policy.

Existing Literature

Private School Vouchers

Just as the MPCP was the first private school choice program to be launched in the United States, it was also the first such program to be evaluated. Several studies assessed the effectiveness of the MPCP in its early years, but all of these analyses can be traced either directly or indirectly to the state's official evaluation, which was authorized

shortly after the program itself began (summarized ultimately in Witte, 2000). Subsequent work analyzed a portion of the data generated by that evaluation and was reported in Greene, Peterson, and Du (1998) and Rouse (1998). These secondary studies differ between themselves and the original Witte (2000) evaluation in their methodological approaches and, as a result, in the outcomes they report. Based on regression models strengthened with Heckman selection corrections, the Witte work generally found no systematic differences in student achievement outcomes between voucher and public school students. The Greene et al. (1998) analysis was based on a subset of voucher participants who had won their voucher via a lottery system to a small number of oversubscribed private schools, finding positive achievement impacts associated with participation. Splitting these differences, Rouse employed a series of approaches including student fixed effects and instrumental variable designs, finding no effect in reading but positive achievement impacts in mathematics for voucher students.

These differences between findings of positive and no achievement effects have subsequently been reflected in the broader school voucher literature. Howell, Wolf, Campbell, and Peterson's (2002) and Howell, Peterson, Wolf, and Campbell's (2006) primary analysis of lottery-based privately funded voucher programs in New York City, Washington, D.C., and Dayton, Ohio demonstrated positive student achievement outcomes for African Americans but not for the overall sample, a finding both confirmed (Barnard, Frangakis, Hill, & Rubin, 2003) and questioned (Krueger & Zhu, 2004) by subsequent analyses of the New York data. Studies of the Cleveland Scholarship Program have similarly found both positive impacts for voucher students (Peterson, Howell, & Greene, 1999) and no significant differences (Metcalf, West, Legan, Paul, & Boone, 2003) depending on the study design and sample, while a recent regression discontinuity analysis by Figlio (2011) on the statewide voucher system in Florida has found small but positive impacts on student achievement, particularly in reading. A recent evaluation of the first federally funded voucher program in Washington, D.C., which was based on a randomized lottery design, reported gains in reading that were at

least marginally statistically significant in years 2 to 4 but found no significant math impacts. That study also showed voucher use to produce substantial gains in student graduation rates (Wolf et al., 2013).

To date, evidence generated on the effects of private school vouchers has come from programs not subject to formal accountability policies. Recent years have brought the initiation of several new voucher programs—or expansion of existing ones—many of which are subject to state testing and accountability policies. Evaluation of several of these programs is underway, and it remains to be seen whether the results of these evaluations mirror those of earlier programs, which were not subject to testing or broader accountability requirements.

High-Stakes Testing

For the public sector, high-stakes testing regimes have recently received considerable attention by education scholars. Although researchers continue to debate the extent to which these initiatives may hold long-lasting promise for educational reform, the evidence suggests that these programs can have positive academic impacts, at least in the short term. Several studies of general accountability policies have shown their implementation to result in student test score gains (e.g., Carnoy & Loeb, 2002; Dee & Jacob, 2011; Hanushek & Raymond, 2005). Similar results have emerged from studies that analyze a specific testing policy affecting a given state or city. For example, a series of articles have found that a school grading policy in Florida led to improved performance for students in the worst schools (Chakrabarti, 2007; Chiang, 2009; Figlio & Rouse, 2005; Rouse, Hannaway, Goldhaber, & Figlio, 2007; West & Peterson, 2006), with recent studies providing similar evidence from New York City (Rockoff & Turner, 2010; Winters & Cowen, 2012) and Chicago (Jacob, 2005). Several articles have stressed that such impacts may differ depending on student background (e.g., Dee & Jacob, 2011; Hanushek & Raymond, 2005; Krieg, 2008; Ladd & Lauen, 2010), perhaps at least partly because schools prioritize students with the highest potential for learning gains (e.g., Booher-Jennings, 2005; Neal & Schanzenbach, 2010).

To date, efforts to hold schools publicly accountable for their performance have remained almost entirely within the public sector. Beyond rudimentary demographic data collection on the part of state and federal education agencies—and beyond the sort of program evaluations reviewed above—private schools are typically left to their own devices to monitor and ultimately improve the quality of the educational product they provide. Whether these schools maximize performance on their own, or through participation in a competitive market for educational services, is the subject of intense debate, even among supporters of private school choice (Finn, Hentges, Petrilli, & Winkler, 2009), and there are important theoretical expectations both in favor of and in opposition to the administration of a high-stakes testing system on private schools serving voucher students (Carlson, Cowen, & Fleming, 2013b). The debate has very recently been heightened with the release of a report by the Thomas B. Fordham Institute arguing that choice schools should be held to testing standards more similar to those in traditional public schools (Emerson, 2014). All of which suggests that an analysis of high-stakes testing impacts on Milwaukee's voucher program can move the school choice literature in a new and important direction.

The MPCP

History and Authorizing Legislation for New Evaluation

When the first official evaluation of the MPCP began in 1990, there were 341 voucher students enrolled in seven secular private schools (Witte, 2000). After that evaluation ended in 1995, the state of Wisconsin expanded the voucher program to include religious schools, and student enrollment in the program grew dramatically as a result. The original Witte (2000), Greene et al. (1998), and Rouse (1998) studies noted above were all based on the pre-expansion data, while during the subsequent period of rapid growth in the MPCP there were no official or scholarly evaluations of the program beyond the state's administrative monitoring. Figure 1 provides a summary of the program's growth—in terms of student enrollment—over its 20-plus year history.

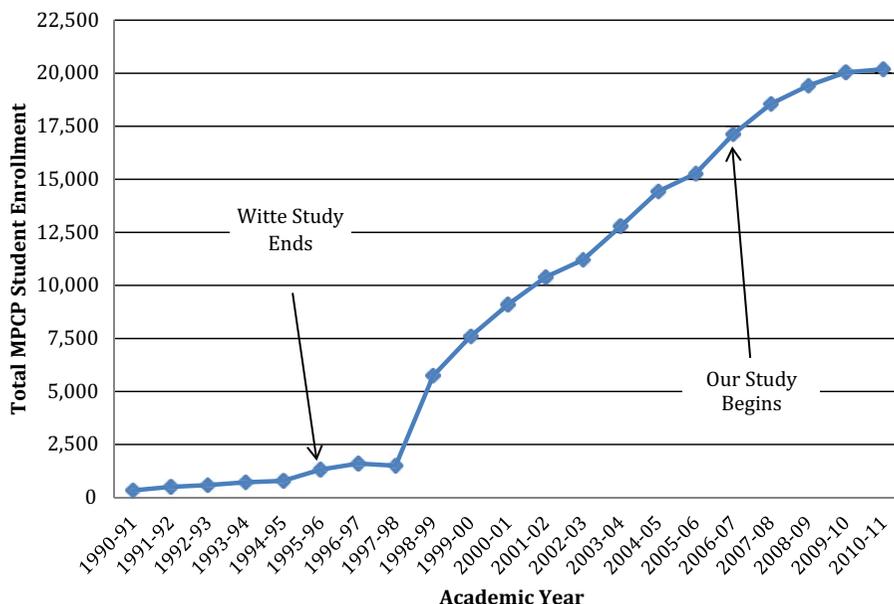


FIGURE 1. MPCP enrollment growth between two state-mandated evaluations.
 Note. MPCP = Milwaukee Parental Choice Program.

In 2005, the state legislature passed Wisconsin Act 125, which not only re-authorized the MPCP but also made several important changes to the program. Before the re-authorization, the total number of students who could participate in the voucher program was limited to 15% of the number of students enrolled in the Milwaukee Public School (MPS) District, which translated to about 15,000 students. Act 125 raised the cap to a nominal figure of 22,500 students. Subject to this cap, any student residing in Milwaukee whose family income was below 175% of federal poverty guidelines was eligible for a voucher, whose value had been capped near US\$6,500 for the past decade through the 2013–2015 school year (Witte, Wolf, Cowen, Fleming, & Lucas-McLean, 2008). Finally, and most relevant here, Act 125 directed us to conduct a 5-year study comparing, in the language of the statute, “the [standardized test] scores of a *representative sample* of pupils participating in the program . . . and the scores of a *comparable group* of pupils enrolled in the [Milwaukee public] school district” (2005 Wisconsin Act 125, emphasis added). Beginning in the fall of 2006, the new law required all schools receiving voucher funds to administer the reading and math portions of the WKCE to all students selected for our evaluation

and submit the test scores to our research team for analysis. The 2005 law also required private schools receiving public money via the MPCP to submit to the evaluation team all school-administered standardized test results—the WKCE or any other standardized test—for all voucher students, not just those in our evaluation sample. The private schools, however, were not required to submit to any additional evaluation procedure. Furthermore, the 2005 law did not require us to consider test scores for individual schools, and confidentiality agreements expressly prohibited us from disclosing data on individual schools to either the public or to the state itself.

Initial Evaluation Design

As evidenced by the studies reviewed above, analysts preparing to begin an evaluation of a school voucher program would find themselves with a variety of examples on which to base their research design—at least in principle. In voucher evaluations, as in studies of other policies, randomized field trials are generally considered the “gold standard” for estimating causal effects, but scholars are also accepting other rigorous designs—such as regression discontinuity techniques, within-estimation (i.e., fixed effects), and

matching—when randomization is not feasible. Wisconsin Act 125, however, mandated analysis of a “representative” sample of MPCP students and a “comparable” group of MPS students, thus eliminating many of these designs from the set of feasible options. In effect, Wisconsin Act 125 mandated construction and implementation of an informative observational evaluation of a voucher program in which students had already chosen private schools.²

Our first step in meeting the evaluative mandates of Wisconsin Act 125 involved leveraging an audited file of the population of approximately 18,000 students confirmed to be enrolled in the MPCP as of September 30, 2006, to draw the requisite representative sample of MPCP students that we would track over the 5-year course of inquiry. We focused primarily on Grades 3 through 8 because tests were required in those grades under No Child Left Behind (NCLB) federal legislation and MPS was therefore already testing in those grades. However, we also selected the entire population of 801 MPCP ninth graders to track high school graduation and college enrollment, but data from these students are not relevant to the purposes of this article.³ After constructing this grade-stratified random sample of MPCP participants—consisting of 2,727 students in total—we informed each MPCP school as to which of their students had been selected and worked with them to collect information on students’ demographic characteristics.⁴ With the cooperation of their schools,⁵ we first administered the WKCE to the panel of Grades 3 to 8 MPCP students in November 2006, the same testing period used by MPS to administer the WKCE to their students.⁶

More challenging than drawing a representative sample of MPCP participants was determining how to construct the “comparable” panel of MPS students that the statute required. Ultimately, we designed this comparison sample using public school students that matched MPCP participants on several key observable characteristics. Specifically, we matched MPS students to MPCP students on the basis of grade level, baseline test score, neighborhood of residence, and demographics. For a more detailed exposition of the process used to construct the comparison sample, see Witte et al. (2008). We also drew a random sample—stratified by grade—of students

enrolled in MPS in 2006 to use as a second group against which the achievement gains of MPCP students could potentially be compared. We tracked all public and private school students over the 5-year period—one baseline year and four outcome years—specified in Act 125. In 2009, however—before our evaluation was complete—the Wisconsin legislature acted again, introducing new rules requiring all voucher schools to test their students with the state’s accountability exam.

New Testing Legislation

Wisconsin Act 28 of 2009 introduced a substantial testing component to the Milwaukee voucher program. This regime, which took effect in the 2010–2011 academic year, was far broader in scope than any other evaluation of the MPCP to date. The law required private schools participating in the MPCP to test all voucher students in Grades 3 to 8 and 10—not just those in our evaluation sample—in reading and math with the WKCE. In addition, the schools were required to test voucher students in Grades 4, 8, and 10 in English/language arts, writing, science, and social studies. All test results had to be submitted to the DPI for review and analysis, but DPI is statutorily prohibited from sanctioning or rewarding schools on the basis of the results. In addition to the testing requirements, schools were also required to develop and report individual academic standards for each subject, and all teachers were required to have obtained a college degree unless waived by the state. Yearly instructional hour minimums—1,050 hours for elementary and 1,137 hours for junior high school—were also instituted for all MPCP schools. The new law also stipulated that, effective immediately in the 2009–2010 school year, schools were required to pay additional administrative fees to the DPI, which oversees the program, and required new schools to obtain accreditation to participate.

Although no explicit sanctions are tied to school assessment results, schools that enroll voucher students are required to participate in the new testing program. Because the vast majority of schools participating in the MPCP serve primarily voucher students (McShane, Kisida, Jensen, & Wolf, 2012), this implies that for most schools, failure to participate could result in

closure. In addition, the public reporting of school test scores implies that, for the first time, Milwaukee parents are able to observe the academic outcomes of individual private schools before deciding which school to attend via the voucher. Such an environment dramatically raises the academic stakes faced by private schools serving voucher students. From the onset of the new testing regime in the fall of 2010, test score results have been reported by DPI by individually named private schools and subsequently analyzed by outside groups in the state (e.g., Dickman & Schmidt, 2012). Carlson et al. (2013b) introduces the possibility that such consequences had an impact on student outcomes, and consider the implications of the policy for outcomes in a variety of contexts in which private agencies provide publicly funded services.

As described above, 2005 Wisconsin Act 125 tasked us with determining the difference between the achievement gains of a “representative” sample of MPCP students and a “comparable” panel of MPS students. We were, in the terms of the voucher literature, tasked with identifying the effect of attending a voucher school, relative to attending a traditional public school. For the representative panel of MPCP students on which we were focusing, the conditions associated with voucher usage between 2006 and 2009 were materially unchanged. Students attended private schools that, beyond permitting us access to those drawn randomly into our panel, were subject to no additional reporting requirements. Our results could and did remain anonymous with respect to the identity of the individual schools involved. The 2009 legislation fundamentally changed these research conditions. Although this change complicated our ability to draw inferences about the causal effect of attending an MPCP school on student test scores, it allowed us to ask a different and perhaps more timely policy question: To what extent do voucher schools respond to high-stakes testing pressures?

Data and Analytic Sample

One fundamental difficulty with estimating the effects of a new high-stakes testing policy is that, often by definition, outcome data may not

be available prior to the new regime. Our ongoing evaluation of the MPCP alleviated this problem, and our subsequent analytical strategy is directly tied to the timing of the new testing law. Figure 2 depicts the data collection process of the evaluation alongside the testing changes that we consider. Critically, this process began several years before the testing legislation was passed in 2009, and before the law took effect in the 2010–2011 school year.

In our initial evaluation of the MPCP, the “treatment” of interest was defined as the sector—MPCP or MPS—in which a student was enrolled in the baseline year of 2006–2007. Such a definition was necessary to preserve the integrity of the matched design outlined in Witte et al. (2008). For our purposes here, where our objective is to identify the high-stakes testing effect, we do not need to rely on this observational match, but do require observations on students that were and were not exposed to the high-stakes testing reform. As such, we prioritize the sector in which students were consistently located during the 3 years in question—2008–2009 to 2010–2011—when defining our analytic sample.⁷ Consequently, our sample contains all students who were either confirmed in MPCP from 2008–2009 to 2010–2011 or confirmed to be in MPS during those three school years, regardless of the sector in which they were originally observed in the baseline year of 2006–2007, and regardless of how the MPS students entered the panel (i.e., via either the original Witte et al., 2008, match or via the construction of the representative, randomly sampled MPS sample). We chose the 3-year period from 2008–2009 to 2010–2011 because it provided a long enough time period to perform the analysis and necessary specification checks over a stable analytic sample.

The timeline of the data collection for the initial evaluation, coupled with multiple features of the educational environment in Wisconsin and demands of our analytical strategy, results in a sample with a particular set of characteristics. First, because our analytic strategy requires test scores from consecutive years, our sample excludes students who were in 10th grade in the first high-stakes year of 2010 as they were not tested as 9th graders the previous year. Second, given that data collection for the initial 5-year

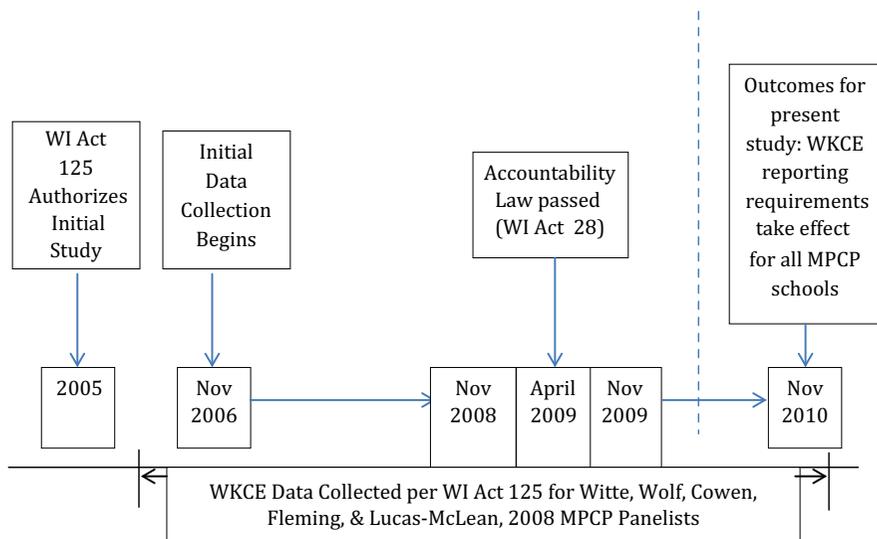


FIGURE 2. Pre- and post-accountability data collection for MPCP panelists.

Note. MPCP = Milwaukee Parental Choice Program; WI = Wisconsin; WKCE = Wisconsin Knowledge and Concepts Examination.

evaluation of the MPCP began in 2006, students who remained in a tested grade in 2010—the first high-stakes year—were primarily in 3rd and 4th grade when data collection began 5 years earlier, although there are a small number of students who were 5th graders in 2006 in our analytic sample; these students were likely retained in grade at some point over the data collection process. As a result, our analytic sample contains primarily students enrolled in Grades 7 and 8 in the first high-stakes testing year of 2010.

Although this sample construction may limit our ability to generalize to a more representative population of MPCP students (i.e., one that includes students in other grades or those who spend a more limited amount of time in the voucher program), it is necessary to cleanly discern which students were and were not subject to the high-stakes reform. Other studies (e.g., Neal & Schanzenbach, 2010) have had to make similar trade-offs when constructing samples for identifying accountability impacts, and among our robustness checks below, we provide alternative formulations of the treatment and comparison groups to ensure that the impact we estimate is not an artifact of sample construction.

Table 1 presents comparisons of student demographics for the students in our sample. Although the original matched design of the

TABLE 1

Descriptive Analytic Sample Statistics

	MPCP	MPS	Difference
Black	.51	.62	-.11**
Hispanic	.37	.24	.13**
Asian	.02	.03	-.01
Native	.00	.01	-.01*
White	.10	.10	<.00
Female	.57	.56	.01
Baseline Math	-.26	-.12	-.15*
Baseline Reading	-.18	-.10	-.07
Grade = 5	.13	.12	.01
Grade = 6	.26	.23	.03*
Grade = 7	.33	.32	.01
Grade = 8	.27	.32	-.05**
Unique <i>n</i>	437	792	

Note. All cells but math and reading are proportions, where significance tests are difference-of-proportions based on approximate normal distribution. Math and reading values are mean Wisconsin Knowledge and Concepts Examination scores standardized by grade against MPS citywide average. MPCP = Milwaukee Parental Choice Program; MPS = Milwaukee Public School.

*Difference significant at $p < .05$. **Difference significant at $p < .01$.

evaluation produced a sample of MPCP students that were statistically indistinguishable from the sample of MPS students on observable

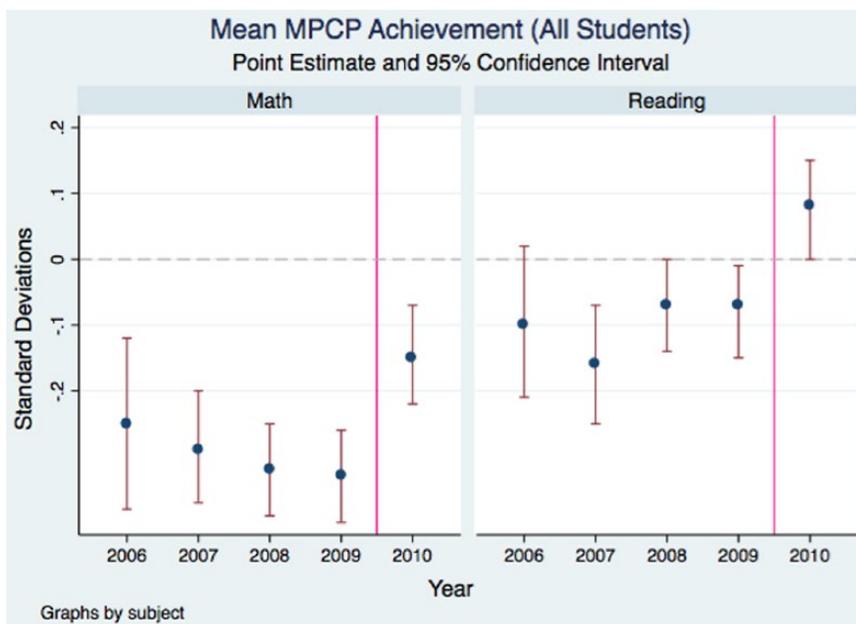


FIGURE 3. Mean MPCP achievement (all students).
 Note. MPCP = Milwaukee Parental Choice Program.

characteristics, our requirement that students maintain the same sector status over the 3-year period—coupled with the inclusion of MPS students selected via the random sample described earlier—yields very different groups of students. Relative to the MPS students in our sample, students confirmed in the MPCP from 2008–2009 to 2010–2011 had similar reading scores in the year they entered the study, but were well behind in math. Note that these scores are standardized against the citywide mean and standard deviation of scores, indicating that the average MPCP and MPS students in our sample are below the citywide average. The fact that the academic performance of the average MPS student in our sample falls below the citywide mean is attributable to the procedure used in the original evaluation that matched MPCP students—who were disproportionately low-performing at baseline—to similarly low-performing MPS students. Table 1 also indicates that MPCP students were far less likely to be African American, far more likely to be Hispanic (and English language learners), and neither more nor less likely to be female than MPS students in our sample. In the models below, we control for differences associated with these observed student characteristics.⁸

Analytic Strategy and Results

The first evidence that the high-stakes testing law affected voucher student performance is depicted pictorially in Figures 3 and 4, which show average student achievement levels in the MPCP, and the average difference between the MPCP and MPS, respectively. The figures make clear that substantial growth in both math and reading occurred among MPCP students in 2010–2011, the first year that the high-stakes testing policy took effect.

Primary Models

To consider these gains further, we rely on the underlying identification assumption that the only systematic difference between the earlier years of our panel and 2010–2011 was the imposition of the new testing regime in the voucher-serving schools. Although we cannot test this assumption directly, we simply have no evidence of any other substantial policy change taking effect in that year, and the reform we have described was indeed so structural in nature—such a break with past private school operations—that we consider it the straightforward interpretation of the educational climate

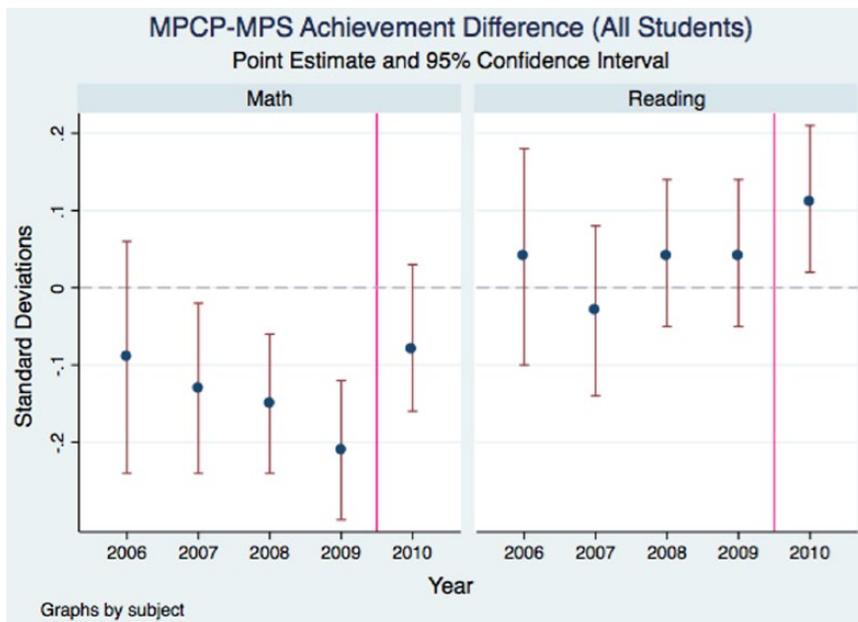


FIGURE 4. MPCP–MPS achievement difference (all students).
 Note. MPCP = Milwaukee Parental Choice Program; MPS = Milwaukee Public School.

unique to 2010–2011. Building off of Carlson et al. (2013b), we use the MPS panel described above as essentially a control group which, although subject to its own public sector testing policy, saw no change in those policies concurrent to the new requirements directed at voucher schools. Use of MPS as the control group is in the same spirit as several other studies (e.g., Dee & Jacob, 2011; Dee, Jacob, & Schwartz, 2013; Grissom, Nicholson-Crotty, & Harrington, 2014) that identify high-stakes reform impacts vis-à-vis jurisdictions (in those cases, states) that had implemented similar reforms earlier and were thus plausibly unaffected by the new law.

With this leverage, we estimate,

$$Y_{it} = \alpha + \delta_{it}(\text{MPCP}_i \times 2010_t) + \beta_1 \text{MPCP}_i + \beta_2 Y_{i,t-2,t-3,t-4} + \beta_3 2010_t + \beta_4 \mathbf{X}_i + \varepsilon_{it}, \quad (1)$$

over all observations in our sample from academic years 2009–2010 and 2010–2011. In this model of math or reading achievement Y , our coefficient of interest, δ_{it} , is a difference-in-differences estimate of the voucher-based high-stakes testing program, not testing per se; β_1 captures 2009–2010 outcome differences between MPCP and MPS; β_2 represents both math and

reading achievement in the pre-high-stakes years of 2008, 2007, and 2006 (the last pre-reform year of 2009 is implicit in the model); β_3 is a time trend shared by both sectors, and \mathbf{X} represents a vector of student race and gender indicators.

The first two columns of Table 2 present the results of estimating Equation 1. The high-stakes estimates are both statistically significant and substantively meaningful, at one tenth of a standard deviation for reading and more than two tenths of a standard deviation in math. These results provide a strong indication of substantial growth in MPCP test scores during the first year the high-stakes testing policy was in place. Like other analysts of similar reform impacts (e.g., Neal & Schanzenbach, 2010), our preferred model requires a particular group of students for whom more than 1 year of test scores is available. To ensure that our estimates of δ_{it} are not somehow contaminated by the requirement that our sample members have five consecutive years of test score data, we estimate the difference-in-differences model without any prior scores (i.e., only requiring 2009 and 2010 scores in the same subject), both for the same sample as our primary models (Columns 3 and 4), and for the unrestricted sample of any

TABLE 2

Estimated Difference-in-Differences High-Stakes Impacts

	Full model		No prior w/full model sample		No prior on all observations	
	Reading	Math	Reading	Math	Reading	Math
δ	.104* (.049)	.252** (.057)	.122 (.064)	.260** (.076)	.108 (.065)	.245** (.073)
2010 dummy	.066* (.028)	.005 (.031)	.054 (.039)	-.018 (.043)	.061 (.038)	-.022 (.042)
MPCP	.058 (.035)	-.092* (.041)	-.075 (.080)	-.322** (.077)	-.015 (.078)	-.277** (.075)
Reading 2006	.122** (.024)	.019 (.025)				
Reading 2007	.220** (.026)	.066* (.027)				
Reading 2008	.398** (.028)	.119** (.031)				
Math 2006	.002 (.027)	.113** (.028)				
Math 2007	.010 (.023)	.148** (.031)				
Math 2008	.136** (.025)	.435** (.032)				
Native	-.023 (.104)	-.155 (.102)	-.846** (.245)	-.899** (.232)	-.766** (.240)	-.889** (.218)
Asian	.045 (.055)	.206** (.059)	-.126 (.125)	.140 (.096)	-.191 (.131)	.044 (.118)
Black	-.086* (.039)	-.083* (.040)	-.667** (.083)	-.710** (.091)	-.687** (.087)	-.736** (.092)
Hispanic	-.019 (.040)	-.003 (.040)	-.291** (.083)	-.249** (.086)	-.310** (.085)	-.271** (.082)
Female	.072** (.027)	.000 (.027)	.239** (.051)	.060 (.050)	.258** (.051)	.092 (.052)
Constant	.263 (.162)	.317* (.151)	.229 (.199)	.276 (.167)	.186 (.177)	.349* (.163)
Observations	2,456	2,454	2,456	2,454	2,656	2,656
R^2	.677	.660	.089	.100	.092	.097

Note. Robust standard errors clustered by school in parentheses. Models also include grade fixed effects. Per Equation 1, δ represents the interaction between outcome year indicator and MPCP flag. MPCP = Milwaukee Parental Choice Program.

* $p < .05$. ** $p < .01$.

students with 2009 or 2010 scores (Columns 5 and 6). As the results indicate, the estimate of the high-stakes testing impact in 2010 remains comparable with those in the preferred specification. We further explore other sample-related comparison issues in other robustness checks below.

Sub-Group Analysis

The accountability studies that we review above have generally—although by no means uniformly—returned evidence of positive high-stakes impacts. Consequently, our findings that the introduction of a similar high-stakes regime

drove test scores upward in the first post-policy year should not be surprising just because the students in our sample attended private rather than public schools. However, as we also note, positive high-stakes impacts have been shown to vary along a variety of academic and demographic characteristics (e.g., Dee & Jacob, 2011; Hanushek & Raymond, 2005; Krieg, 2008; Ladd & Lauen, 2010). Moreover, previous studies of school vouchers have noted differential voucher impacts based on student background characteristics as well (e.g., Howell et al., 2006; Wolf et al., 2013).

To begin to consider the issue here, we estimate Equation 1 separately by quartile of 2006 test scores—the first scores we collected in the initial evaluation. Table 3 provides these results, which indicate strong high-stakes testing impacts for math in all but the lowest quartile. On one hand, this suggests that the clear testing impacts for math in the full sample results (Table 2) were particularly positive for students with higher initial ability. For reading, on the other hand, only the lowest quartile approaches statistical significance (at $p < .10$). If there is somewhat conflicting evidence between the subjects for the lowest initial performers, the broader pattern in Table 3 suggests that the high-stakes testing impacts are evident for different initial ability groups, with no single quartile appearing to drive the full sample results.

More nuance is apparent in Table 4, which presents results from the estimation of Equation 1 separately for students by race and gender. For the most part, the high-stakes impacts observed in Table 2 are consistent across sub-groups, particularly in the sense that the impact appears weaker for reading. The effects for Blacks are positive and significant for both subjects, while only math is positive for Hispanics and White or Asian students. The point estimates for males are lower in both subjects than for females, although the math effects for males are comparable.

Additional Specifications and Robustness Tests

Falsification Test

Although the high-stakes reform was passed in 2009, the primary testing requirements did not take place until 2010, making this the first reform year and the one of central interest above. To

check that this year was truly different from the others, we provide a falsification or “placebo” test in which we estimate a model analogous to Equation 1 over all sample observations in 2008–2009 and 2009–2010—the last two pre-reform years. In this model, we replace the 2010 indicator with a 2009 indicator and interact the 2009 dummy with the MPCP flag. If the difference-in-differences results we observed in 2010 were driven by, for example, an upward trend induced by some other intervention targeted to the MPCP—or if the achievement of MPCP students was simply on a different growth trajectory from that of the MPS students in our sample—our difference-in-differences term in Equation 1 may be capturing that confounding trend instead of the high-stakes impact. Figures 3 and 4 suggest this is not the case, and Table 5 provides further confirmation. In both reading and math, the coefficients on the placebo difference-in-differences terms are both statistically insignificant and substantively small.⁹

Within-Student MPCP Differences

Our primary identification strategy relies on a comparison with MPS students, who were not exposed to a new high-stakes reform law in the period that we examine. Figures 2 and 3 imply that the MPCP growth we estimate in the post-reform year was meaningfully different than in the pre-reform years, but the nature of our comparison sample renders it conceivable that our estimates could overstate the growth actually occurring for MPCP students. As a robustness check against the difference-in-differences framework, we estimate a model that compares MPCP students’ 2010 scores to their achievement levels in previous years. This interrupted time-series model, which contains student fixed effects, is estimated over the observations from our sample of MPCP students—MPS students are excluded—in 2008, 2009, and 2010:

$$Y_{it} = 2010_i \alpha_1 + 2008_i \alpha_2 + \phi_i + \varepsilon_{it}, \quad (2)$$

where 2009 serves as the omitted year. Consequently, a positive estimate of the 2010 year indicator implies that relative to the last pre-reform year, the average level of student achievement is significantly higher. The base year in Equation 1 for the difference-in-differences comparison is also 2009, which means that a positive

TABLE 3

High-Stakes Impacts by Initial Achievement Sub-Group

2006 quartile	0%–25%		26%–50%		51%–75%		76%–100%	
	Reading	Math	Reading	Math	Reading	Math	Reading	Math
δ	.233 (.129)	.105 (.122)	.085 (.106)	.284** (.085)	.010 (.058)	.328** (.081)	.085 (.071)	.331** (.089)
2010 dummy	.084 (.070)	.073 (.074)	.114* (.044)	–.050 (.057)	.041 (.037)	.016 (.029)	.053 (.038)	.017 (.043)
MPCP	.040 (.108)	–.065 (.092)	.043 (.070)	–.085 (.063)	.107 (.057)	–.148** (.056)	.076 (.054)	–.122 (.068)
Reading 2006	.091 (.048)	.055 (.050)	.404** (.123)	–.039 (.041)	.350* (.177)	.058 (.066)	.047 (.076)	–.102* (.049)
Reading 2007	.173** (.042)	.127* (.049)	.174** (.046)	–.018 (.034)	.254** (.051)	–.005 (.048)	.437** (.054)	.107 (.055)
Reading 2008	.277** (.050)	.118* (.055)	.451** (.053)	.119* (.047)	.438** (.058)	.107** (.037)	.416** (.057)	.118 (.063)
Math 2006	.063 (.044)	.113 (.058)	–.014 (.043)	.339* (.132)	–.065 (.043)	.213 (.118)	–.071 (.037)	.066 (.053)
Math 2007	.017 (.045)	.103 (.064)	–.004 (.035)	.223** (.048)	.036 (.062)	.111* (.045)	–.013 (.048)	.133** (.047)
Math 2008	.178** (.040)	.351** (.064)	.079 (.055)	.379** (.048)	.120* (.057)	.517** (.053)	.111* (.054)	.616** (.055)
Native	.002 (.143)	–.236 (.220)	.087 (.204)	–.195 (.264)	–.018 (.198)	–.064 (.129)	NA	–.421** (.080)
Asian	.103 (.161)	.485* (.231)	.010 (.100)	.375** (.100)	.157 (.148)	–.009 (.120)	.087 (.094)	.157 (.097)
Black	–.043 (.102)	–.057 (.198)	–.092 (.099)	.006 (.069)	.042 (.062)	–.168** (.062)	–.133* (.064)	–.059 (.061)
Hispanic	.006 (.123)	.012 (.198)	.006 (.100)	.067 (.076)	.091 (.059)	–.046 (.064)	–.029 (.055)	–.061 (.060)
Female	.149* (.068)	.008 (.065)	.061 (.048)	.061 (.048)	.014 (.042)	–.013 (.047)	.054 (.034)	–.031 (.043)
Constant	.006 (.244)	.245 (.266)	.445** (.143)	.402 (.216)	.902** (.161)	.509** (.077)	–.233* (.091)	–.148 (.194)
Observations	585	588	615	601	632	626	624	639
R^2	.407	.345	.416	.438	.505	.510	.598	.660

Note. Robust standard errors clustered by school in parenthesis. Models also include grade fixed effects and student race/gender demographics. Columns report estimates divided into quartiles of 2006 (baseline) averaged math and reading WKCE scores. MPCP = Milwaukee Parental Choice Program; WKCE = Wisconsin Knowledge and Concepts Examination.

* $p < .05$. ** $p < .01$.

estimate of α_1 provides strong evidence that the differences estimated in Equation 1 are not driven by our comparison with MPS. Indeed, as Table 6 suggests, the MPCP outcomes in 2010 were markedly higher relative to earlier years for the same students.

“Intent to Treat” and Representative MPS Comparisons

Previous work has demonstrated that schools often respond strategically to the implementation of a performance measurement system (e.g., Cullen & Reback, 2006; Figlio & Getzler, 2002;

TABLE 4

High-Stakes Impacts by Race and Gender

	Black		Hispanic		White or Asian		Females		Males	
	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math
δ	.163*	.182*	.108	.311**	-.077	.180*	.164**	.279**	.015	.215*
	(.074)	(.079)	(.077)	(.093)	(.086)	(.084)	(.058)	(.069)	(.076)	(.083)
2010 dummy	.050	-.041	.088*	.021	.091	.144*	.031	.000	.097*	.013
	(.040)	(.046)	(.042)	(.039)	(.054)	(.058)	(.032)	(.036)	(.044)	(.044)
MPCP	.091	.020	.004	-.200**	.076	-.124	.051	-.104	.084	-.072
	(.056)	(.059)	(.048)	(.068)	(.076)	(.068)	(.047)	(.055)	(.062)	(.063)
Reading 2006	.089**	.070*	.208**	-.041	.150**	-.221**	.083*	.010	.148**	.025
	(.030)	(.033)	(.047)	(.050)	(.053)	(.058)	(.036)	(.038)	(.034)	(.038)
Reading 2007	.196**	.032	.191**	.084	.452**	.301**	.267**	.112**	.187**	.030
	(.035)	(.034)	(.054)	(.048)	(.051)	(.076)	(.043)	(.039)	(.032)	(.038)
Reading 2008	.415**	.120**	.355**	.103*	.301**	.079	.427**	.105*	.373**	.131**
	(.038)	(.042)	(.059)	(.049)	(.070)	(.063)	(.041)	(.041)	(.042)	(.046)
Math 2006	.010	.120**	-.000	.094	-.060	.125*	-.001	.091**	.010	.128**
	(.039)	(.037)	(.039)	(.055)	(.069)	(.058)	(.032)	(.033)	(.040)	(.038)
Math 2007	.002	.116**	-.007	.217**	.116	.189**	-.010	.105**	.021	.189**
	(.028)	(.037)	(.040)	(.053)	(.068)	(.065)	(.030)	(.039)	(.036)	(.037)
Math 2008	.162**	.444**	.145**	.393**	-.017	.464**	.111**	.456**	.164**	.418**
	(.033)	(.040)	(.053)	(.060)	(.060)	(.068)	(.032)	(.043)	(.041)	(.043)
Constant	.456**	.341	.119	.206	-.003	.034	.382*	.219	.221	.397*
	(.133)	(.232)	(.219)	(.215)	(.126)	(.106)	(.156)	(.276)	(.203)	(.163)
Observations	1,379	1,378	683	682	379	379	1,317	1,313	1,139	1,141
R ²	.628	.606	.674	.633	.756	.751	.694	.667	.660	.659

Note. Robust standard errors clustered by school in parenthesis. Models include grade fixed effects; race sub-group models also include gender; gender sub-group models also include race. MPCP = Milwaukee Parental Choice Program.

* $p < .05$. ** $p < .01$.

Jacob, 2005; Jacob & Levitt, 2003). One strategy that MPCP schools could potentially employ in response to the new high-stakes testing policy could involve manipulation of the pool of low-performing students who must take the WKCE. Such a scenario is quite possible, especially if schools were particularly concerned about their test scores in the first year (2010–2011) under the policy.

Such a story would fit qualitatively with findings presented in Cowen, Fleming, Witte, and Wolf (2012), which indicates that the lowest performing students are those most likely to leave MPCP at any given time—perhaps because they are not served as well there. More to the point, Carlson, Cowen, and Fleming (2013a) show that these low-performing students exhibited substantial achievement gains—between 0.2 and 0.3 standard deviations—upon moving back to MPS, and relatively few MPS panelists transferred into

the MPCP during the original evaluation. Although these studies were based entirely on pre-reform years in MPCP—thus the lowest performers were leaving MPCP even before the high-stakes testing reform took place—it is possible that the new law strengthened this trend, perhaps in conjunction with the intensified focus on higher performers suggested by our results above. It is also possible that attrition out of MPCP, whether reform-induced or not, implies that the estimated reform impact is somehow an artifact of our analytic sample construction.

The nature of our main parameter of interest in this study—the effect of the private school high-stakes testing policy on students subject to the policy—requires us to analyze students who were consistently located in the same sector over the 3-year period we study. As a consequence of this reality, it is possible that eliminating students who left the MPCP after implementation

TABLE 5
Falsification (Placebo) Tests

	Reading	Math
δ	.055 (.042)	.015 (.049)
2009 dummy	-.012 (.027)	.019 (.025)
MPCP	-.008 (.042)	-.165** (.046)
Reading 2006	.219** (.028)	.010 (.026)
Reading 2007	.437** (.027)	.188** (.032)
Math 2006	.045* (.020)	.209** (.025)
Math 2007	.156** (.021)	.447** (.024)
Native	.129 (.150)	-.046 (.089)
Asian	-.076 (.049)	.154** (.059)
Black	-.034 (.038)	-.041 (.046)
Hispanic	-.010 (.037)	.024 (.043)
Female	.104** (.028)	.013 (.023)
Constant	.072 (.147)	.029 (.170)
Observations	2,939	2,939
R^2	.638	.616

Note. Robust standard errors clustered by school in parenthesis. Models also include grade fixed effects. Per Equation 1, δ represents the interaction between outcome year indicator and MPCP flag, where here 2009 is outcome year. MPCP = Milwaukee Parental Choice Program.
* $p < .05$. ** $p < .01$.

of the testing policy from our analytic sample resulted in the exclusion of low-achieving students that, in this scenario, MPCP schools strategically counseled out. As a first step in assessing the plausibility of this scenario, we simply compared the prior-year achievement outcomes of students who exited the MPCP in 2010–2011 with the prior-year achievement outcomes of students who exited the MPCP in earlier years.¹⁰ The results reveal no differences between the

TABLE 6
Within-Student Models (MPCP Only)

	Reading	Math
2010 dummy	.154** (.045)	.241** (.041)
2008 dummy	-.046 (.035)	-.062 (.036)
Constant	.019 (.019)	-.202** (.022)
Observations	846	844
R^2	.056	.121

Note. Robust standard errors clustered by school in parenthesis. Models include student fixed effects; the last pre-reform year of 2009 is the reference category for 2010 and 2008 indicator variables. MPCP = Milwaukee Parental Choice Program.
* $p < .05$. ** $p < .01$.

prior-year achievement of students who exited the MPCP in 2010–2011 and those who exited in earlier years.

A simple test borrowed from the randomized control trial framework can help provide further evidence on this issue. For this test, we estimate Equation 1 over an analytic sample where the MPCP, or “treatment,” group is now defined as all student-year observations from 2008–2009 to 2010–2011 for all students enrolled in MPCP in 2008–2009, not just those who remained in the voucher sector all 3 years. In effect, we are estimating an intention-to-treat (ITT) parameter as opposed to the treatment-on-the-treated (TOT) parameters we estimated above.¹¹ If selective attrition were responsible for the observed achievement increases, then we would expect the ITT estimates to be smaller than the TOT estimates. The first two columns of Table 7 provide the ITT parameter estimates. For both reading and math, these results are strikingly similar to those in Table 2. Although not fully definitive, this is strong evidence that the achievement gains in voucher schools were not driven by exits of lower performing students from the voucher program—whether these exits were policy-induced or otherwise systematic with respect to our outcomes.¹²

Finally, the last two columns in Table 7 provide one more check to ensure that our analytic sample construction did not create a misleading

TABLE 7
 “ITT” and Representative MPS Comparisons

	ITT		Baseline MPCP vs. representative MPS panel	
	Reading	Math	Reading	Math
δ	.100** (.038)	.237** (.052)	.086 (.050)	.256** (.064)
2010 dummy	.045 (.025)	-.007 (.033)	.066* (.030)	-.012 (.038)
MPCP	.067* (.026)	-.051 (.033)	.087* (.039)	-.056 (.053)
Reading 2006	.121** (.021)	.015 (.023)	.166** (.034)	.001 (.034)
Reading 2007	.217** (.024)	.068** (.021)	.195** (.041)	.073 (.042)
Reading 2008	.385** (.023)	.116** (.025)	.429** (.039)	.134** (.035)
Math 2006	-.004 (.024)	.116** (.025)	-.034 (.036)	.112** (.040)
Math 2007	.025 (.022)	.158** (.029)	.075* (.033)	.198** (.037)
Math 2008	.137** (.023)	.429** (.031)	.100** (.033)	.416** (.038)
Native	-.035 (.088)	-.243* (.107)	.112 (.129)	.128 (.072)
Asian	.080 (.059)	.214** (.062)	.055 (.063)	.217** (.063)
Black	-.080* (.036)	-.082* (.036)	-.065 (.051)	-.097 (.053)
Hispanic	-.006 (.039)	-.008 (.037)	-.018 (.049)	-.015 (.049)
Female	.089** (.027)	.001 (.025)	.056 (.037)	.048 (.036)
Constant	.200 (.131)	.271* (.126)	.131 (.265)	.389 (.278)
Observations	2,933	2,929	1,395	1,393
R^2	.669	.658	.682	.670

Note. Robust standard errors clustered by school in parenthesis. Models include grade fixed effects and student race/gender demographics. In ITT models, MPCP students defined by status in 2008. ITT = intention-to-treat; MPCP = Milwaukee Parental Choice Program.

* $p < .05$. ** $p < .01$.

comparison. Recall that the initial evaluation included a representative panel of grade-stratified, randomly sampled MPS students that the Witte et al. (2008) evaluators intended to track as a second possible comparison group for the panel of MPCP

students. Constructing the analytic sample here, it did not matter how an MPS student was drawn into the sample—either by the Witte et al. match, or via the random sampling, or for that matter whether they were originally a voucher student

who had transferred to MPS by 2008. All we needed was a group of students who were in the same sector well before and after the high-stakes reform. Our primary results tell us how students in the MPCP were affected by the high-stakes testing reform relative to students who were unexposed to the new law by virtue of their location in MPS for any reason. These results do not tell us how the reform affected MPCP students relative to a typical MPS student. The final two columns in Table 7 provide such evidence, where the comparison is directly between MPCP students exposed to the high-stakes reform and the representative panel of MPS students drawn by Witte et al. The results remain very similar to those from the models above.

Discussion

This study began as a state-mandated evaluation of the MPCP, the nation's oldest and largest urban voucher program. As we have described, the legislation authorizing this study mandated that we compare the achievement growth of MPCP students to that of a comparable group of students in public schools over a period of 5 years. Prior to the final year of data collection, however, the Wisconsin state legislature instituted a substantial high-stakes testing requirement for private schools receiving voucher funds. While schools were previously required to provide test scores for our study's panelists only, the new law required schools to test all voucher students and report the results to the state; while our study was prohibited from disaggregating our results by private school, the new testing system resulted in the public reporting of each private school's scores for the first time. This law fundamentally changed the conditions in which our initial study had been established, but our collection of data both before and after the law allowed us to ask a different and timely policy question: whether schools receiving voucher funds could realize productivity gains from high-stakes testing pressures.

Our study is among the first empirical analyses capable of identifying the effect of adding a test-based accountability element to a private school voucher program. In its emphasis on standardized test administration and public reporting of results, the policy we analyze contains many of the same features of the testing

system imposed on public schools under NCLB. We use difference-in-differences techniques to estimate the effect of the reform on the achievement of voucher students against a comparison group of students attending MPS who received no new policy change. Our general results indicate that introduction of the testing policy resulted in substantial student achievement gains, particularly in math, and in that subject perhaps for students with higher levels of initial ability. This may simply indicate that as new legislation required each school's outcomes to be measured and, critically, to be made public, schools simply worked to maximize their aggregate scores. It is also possible that the newly high-stakes nature of the testing induced the private voucher schools to perform more effective test preparation for all of their students, and the higher performing among the students benefited somewhat more from the preparation.

This evidence has particular salience in the debate over school choice policy in general, and private school voucher programs in particular. Generally, we do not view the implications of the present findings as inherently positive or negative for private school voucher programs, or for the larger set of school choice policies. Indeed, if our results are to be considered as evidence for or against the efficacy of the policies we address in this article, we interpret them as evidence of the potential positive effects that well-designed testing and reporting systems can have on student achievement outcomes in the private sector. In our specific context, it seems likely that the high-stakes policy prompted private schools—as NCLB did their public counterparts—to take actions to avoid low performance ratings. If our results suggest that high-stakes policies may spur improvement among private schools receiving public funds, they do not go so far as to suggest that increased test-based accountability will change the fundamental dynamic of who chooses and who benefits from private school vouchers.

Although our identification strategy and our empirical results are fairly straightforward, we acknowledge that the results of a single study need to be interpreted with caution. Given data limitations, we are unable to examine student achievement growth beyond the first year after implementation of increased accountability. It is possible that the results presented here indicate a

1-year jolt in student achievement, rather than the beginning of sustained growth in test scores under a new high-stakes regime. Future research should examine the long-term effects of such policies. Furthermore, the city of Milwaukee and the MPCP itself may have particular characteristics that limit the generalizability of these results to other contexts. For example, the effects of a similar test-based accountability policy on students and schools in a new or small school voucher program could very well differ from those we identify here, which come from a large, mature voucher program.

More than 100 private schools serve in excess of 20,000 students receiving publicly funded vouchers in Milwaukee—making the program approximately one fifth the size of the surrounding public school district. The educational environment in the city, created in no small part by the voucher program, is one in which private and public schools quite literally compete for students across the city. The positive high-stakes impacts presented above indicate that, in Milwaukee at least, voucher schools had room for improvement above and beyond whatever market forces were at work within this choice-rich educational environment. For students in such urban environments, the good news is that improvement did occur. These outcomes imply that large-scale choice programs can remain at—and even build on—their potential as viable alternatives to traditional public schools when both sectors are held to the same reporting and performance standards.

Authors' Note

The responsibility for the content of the article remains the authors' alone.

Acknowledgments

The authors gratefully acknowledge the cooperation of the Wisconsin State Department of Public Instruction, Milwaukee Public Schools, and the private schools participating in the Milwaukee Parental Choice Program for data access and advice necessary to conduct this study. All errors are our own.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors gratefully acknowledge the research support of the Annie E. Casey, Joyce, Kern Family, Lynde and Harry Bradley, Robertson and Walton Family Foundations.

Notes

1. The states of Maine and Vermont have operated “town tuitioning” programs since the 19th century whereby the government funds the private schooling of students in rural areas that lack public schools. Although some policy analysts classify those initiatives as school voucher programs, we do not because, unlike all other voucher programs in the United States, payments are made directly from the government to the private schools and, therefore, they are restricted to non-sectarian schools.

2. In addition, random assignment comparisons were impossible for a number of reasons, most important of which was that we were not given access to waiting lists or rejected students, and even then, those lists would only be available for some schools and grades that were oversubscribed.

3. For analysis of these outcomes, see Cowen, Fleming, Witte, Wolf, and Kisida (2013).

4. As a safeguard against a loss of power caused by students who reached terminal grades and general study attrition, the initial evaluation also drew refresh samples of 400 to 500 Milwaukee Parental Choice Program (MPCP) third graders in the falls of 2007 and 2008. The specification of our models used to estimate the effect of the accountability policy does not allow these students to be included in the analytic sample used in this article.

5. The research team provided training to the staff of MPCP schools regarding how to administer the Wisconsin Knowledge and Concepts Examination (WKCE). This was to achieve as consistent and valid administration of the test as possible, and was especially geared toward the MPCP schools that were not already using the WKCE. In the first 2 years of the study, 2006 and 2007, the research team provided professional test proctors to some MPCP schools that did not think they were sufficiently prepared to administer the tests themselves.

6. Like other public schools, Milwaukee Public School (MPS) had ample experience with testing on the state exam by the time the initial evaluation occurred. If this experience translated into higher student growth scores relative to students in schools that did not routinely use the exam (for reasons associated with “teaching to the test”), then a public-private comparison could disadvantage voucher students on that outcome measure—at least insofar as the comparison is used to

identify the effect of attending private school. Such a concern does not apply here, where we are estimating the effect of a new accountability policy.

7. As noted in Cowen, Fleming, Witte, and Wolf (2012) and Carlson, Cowen, and Fleming (2013a), there is a substantial attrition of children from the MPCP to MPS each year. In our case, because of the accumulation of yearly MPCP-to-MPS transfers in the five academic years from 2006 to 2010, only 38% of our MPCP students remained in the voucher program every year through 2010. Our estimates below are robust to more generalized panel construction.

8. Additional administrative demographics are problematic here, particularly the typically employed indicators of free/reduced-lunch eligibility, English language learning status, and special academic needs. Simply put, there is no uniform use of these flags throughout the private sector, and although some students might bear one or more of these indicators in MPS, missing data or data indicating the absence of one indicator does not confirm the student is not eligible for such a classification. For example, all students eligible for MPCP fall below 175% of the federal poverty line when they enter the program, even though many schools do not serve or participate in free-lunch programs (Witte, Wolf, Cowen, Fleming, & Lucas-McLean, 2008). Similarly, there is strong evidence that, although special needs students are somewhat less likely to enroll in MPCP, the public-private difference is severely overstated—by as much as 50%—by private schools' lack of either participation in formal special needs programs or an unwillingness to single out individual students with such labels (Wolf, Fleming, & Witte, 2012). Absent a uniform definition of free/reduced lunch, English language learning status, and special needs across schools and sectors, we do not employ them as covariates in our models here, leaving our vector of multi-year prior outcome lags to capture lingering sources of variation in our dependent variables. As a check, we estimated versions of these models with “confirmed” flags for each indicator included, but these do not substantially change our reported results.

9. In a footnote above, we raise the possibility that MPS's experience with the WKCE exam could bias a public-private comparison if we used it to identify the effect of attending private school. In such a scenario, the private school scores could be artificially depressed relative to true academic gains. In our case here, however, such bias would not only have to be induced (or exacerbated) by something in 2010 relative to other years—a possibility ruled out with the placebo test—but would in any case imply that the accountability coefficients we estimate actually understate the policy's impact on private school students.

10. To further illustrate the nature of this test, we compared the 2008–2009 achievement outcomes

for students who left the MPCP in 2010–2011 to the 2007–2008 achievement outcomes for students who left the MPCP in 2008–2009.

11. We thank Tom Dee for suggesting this test during an initial discussion of an earlier draft of this article at the Association for Education Finance and Policy meetings in March 2012.

12. A related possibility is that MPCP schools responded to the impending performance measurement regime not by pushing students away from their schools entirely but by grade retention of low-performers in earlier grades. That response would represent a sort of “middle ground” between no strategic response to accountability and a targeted attempt to dramatically change the testing population by counseling out poor performers, particularly for the many MPCP schools dependent on each voucher student for survival. We have some anecdotal evidence that, generally, some MPCP schools are hesitant to retain children for reasons similar to their hesitation to label children as special needs, described above (Wolf et al., 2012). If true, a response that included new retention efforts would represent a shift from past behavior. In our data, only 1% of MPCP students were ever retained from 2008 to 2010, compared with 5% of students in MPS, and the first year under the policy does not appear to have demonstrated any increase in the probability that a student was retained.

References

- Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association*, 98, 299–311.
- Booher-Jennings, J. (2005). Below the bubble: “Educational triage” and the Texas Accountability System. *American Educational Research Journal*, 42, 231–268.
- Carlson, D. E., Cowen, J. M., & Fleming, D. J. (2013a). Life after vouchers: What happens to private school students when they return to the public sector? *Educational Evaluation and Policy Analysis*, 35, 179–199.
- Carlson, D. E., Cowen, J. M., & Fleming, D. J. (2013b). Third-party governance and performance measurement: A case study of private school vouchers. *Journal of Public Administration Research and Theory*. Advance online publication. doi:10.1093/jopart/mut017
- Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24, 305–331.

- Chakrabarti, R. (2007). *Vouchers, public school response, and the role of incentives: Evidence from Florida* (Federal Reserve Bank of New York: Staff Report No. 37). Retrieved from http://www.newyorkfed.org/research/staff_reports/sr306.pdf
- Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, 93, 1045–1057.
- Cowen, J. M., Fleming, D. J., Witte, J. F., & Wolf, P. J. (2012). Going public: Who leaves a large, longstanding, and widely available urban voucher program. *American Educational Research Journal*, 49, 231–256.
- Cowen, J. M., Fleming, D. J., Witte, J. F., Wolf, P. J., & Kisida, B. (2013). School vouchers and student attainment: Evidence from a state-mandated study of the Milwaukee Parental Choice Program. *Policy Studies Journal*, 41, 147–167.
- Cullen, J. B., & Reback, R. (2006). *Tinkering toward accolades: School gaming under a performance accountability system* (NBER Working Paper 12286). Retrieved from <http://www.nber.org/papers/w12286>
- Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, 30, 418–446.
- Dee, T. S., Jacob, B., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, 35, 252–279.
- Dickman, A., & Schmidt, J. (2012). *Research brief: Significant growth in choice*. Milwaukee, WI: Public Policy Forum.
- Emerson, A. (2014). *Public accountability & private-school choice*. Washington, DC: Thomas B. Fordham Institute. Retrieved from <http://www.edexcellence.net/publications/public-accountability-private-school-choice>
- Figlio, D. N. (2011). *Evaluation of the Florida Tax Credit Scholarship Program participation, compliance and test scores in 2009-10* (Report to the Florida Department of Education). Retrieved from https://www.floridaschoolchoice.org/pdf/FTC_Research_2009-10_report.pdf
- Figlio, D. N., & Getzler, L. S. (2002). *Accountability, ability, and disability: Gaming the system* (NBER Working Paper 9307). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=347079
- Figlio, D. N., & Rouse, C. E. (2005). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics*, 90, 239–255.
- Finn, C. E., Hentges, C. M., Petrilli, M. J., & Winkler, A. M. (2009). *When private schools take public dollars: What's the place of accountability in school choice programs?* Washington, DC: The Thomas B. Fordham Institute.
- Greene, J. P., Peterson, P. E., & Du, J. (1998). Effectiveness of school choice: The Milwaukee voucher experiment. *Education and Urban Society*, 31, 190–213.
- Grissom, J. A., Nicholson-Crotty, S., & Harrington, J. R. (2014). Estimating the effects of No Child Left Behind on teachers' work environments and job attitudes. *Educational Evaluation and Policy Analysis*. Advance online publication. doi:10.3102/0162373714533817
- Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24, 297–327.
- Howell, W. G., Peterson, P. E., Wolf, P. J., & Campbell, D. E. (2006). *The education gap: Vouchers and urban schools*. Washington, DC: Brookings Institution Press.
- Howell, W. G., Wolf, P. J., Campbell, D. E., & Peterson, P. E. (2002). School vouchers and academic performance: Results from three randomized field trials. *Journal of Policy Analysis and Management*, 21, 191–217.
- Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics*, 89, 761–796.
- Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118, 843–877.
- Krieg, J. M. (2008). Are students left behind? The distributional effects of the No Child Left Behind Act. *Education Finance and Policy*, 3, 250–281.
- Krueger, A. B., & Zhu, P. (2004). Another look at the New York City voucher experiment. *American Behavioral Scientist*, 47, 658–698.
- Ladd, H. F., & Lauen, D. S. (2010). Status versus growth: The distributional effect of school accountability policies. *Journal of Policy Analysis and Management*, 29, 426–450.
- McShane, M. Q., Kisida, B., Jensen, L. I., & Wolf, P. J. (2012). *Milwaukee Parental Choice Program: Descriptive report on participating schools*. Fayetteville, AR: University of Arkansas Department of Education Reform. Retrieved from <http://www.uaedreform.org/milwaukee-parental-choice-program-descriptive-report-on-participating-schools-2010-11/>
- Metcalf, K. K., West, S. D., Legan, N. A., Paul, K. M., & Boone, W. J. (2003). *Evaluation of the Cleveland Scholarship and Tutoring Program: Summary Report 1998-2002*. Bloomington: Indiana University.
- Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based

- accountability. *The Review of Economics and Statistics*, 92, 263–283.
- Peterson, P. E., Howell, W. G., & Greene, J. P. (1999). *An evaluation of the Cleveland Voucher Program after two years*. Cambridge, MA: Harvard University Program on Education Policy and Governance. Retrieved from <http://www.hks.harvard.edu/pepg/PDF/Papers/clev2ex.pdf>
- Rockoff, J., & Turner, L. J. (2010). Short run impacts of accountability on school quality. *American Economic Journal: Economic Policy*, 2, 119–147.
- Rouse, C. E. (1998). Private school vouchers and student achievement: An evaluation of the Milwaukee Parental Choice Program. *Quarterly Journal of Economics*, 113, 553–602.
- Rouse, C. E., Hannway, J., Goldhaber, D., & Figlio, D. (2007). *Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure* (CEPS Working Paper No. 156). Retrieved from <http://www.nber.org/papers/w13681>
- West, M. R., & Peterson, P. E. (2006). The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments. *Economic Journal*, 116, 46–62.
- Winters, M. A., & Cowen, J. M. (2012). Grading New York: Accountability and student proficiency in American's largest school district. *Educational Evaluation and Policy Analysis*, 34, 313–327.
- Witte, J. F. (2000). *The market approach to education: An analysis of America's First Voucher Program*. Princeton, NJ: Princeton University Press.
- Witte, J. F., Wolf, P. J., Cowen, J. M., Fleming, D. J., & Lucas-McLean, J. (2008). *MPCP longitudinal educational growth study baseline report* (University of Arkansas Education Working Paper Archive). Retrieved from http://www.uark.edu/ua/der/EWPA/Research/School_Choice/1806.html
- Wolf, P. J., Fleming, D. J., & Witte, J. F. (2012). Do voucher schools serve students with disabilities? *Education Next*, 12(3). Retrieved from <http://educationnext.org/special-choices/>
- Wolf, P. J., Kisida, B., Guttman, B., Puma, M., Eissa, N., & Rizzo, L. (2013). School vouchers and student outcomes: Experimental evidence from Washington, D.C. *Journal of Policy Analysis and Management*, 32, 246–270.

Authors

JOHN F. WITTE is a professor emeritus of political science at the University of Wisconsin–Madison. His research interests include public policy analysis and process, with specialties in education and tax policy and politics. His current focus is on charter schools, open enrollment, and Milwaukee's voucher program. During the 2012–2013 academic year, he was founding dean of the School of Humanities and Social Sciences, Nazarbayev University, Astana, Kazakhstan.

PATRICK J. WOLF is a professor of education policy, 21st century endowed chair in school choice, and distinguished professor at the University of Arkansas in Fayetteville. His research interests include school vouchers, charter schools, special education, and public administration. His current projects include school voucher evaluations in Indiana, Louisiana, and India as well as a national study of charter school financing.

JOSHUA M. COWEN (corresponding author) is an associate professor of educational policy in the College of Education at Michigan State University. His research concerns school choice and mobility, teacher quality, and policy reform. Recent work has been published in *Educational Researcher*, *Education Finance and Policy*, and *Teachers College Record*. He can be contacted at jcowen@msu.edu.

DEVEN E. CARLSON is an assistant professor in the political science department at the University of Oklahoma. His research agenda explores the operations of public policies and analyzes their effects on political, social, and economic outcomes of interest.

DAVID J. FLEMING is an assistant professor in the political science department at Furman University. His research interests include civic education, policy feedback, school choice, and Montessori education.

Manuscript received May 8, 2013

First revision received September 24, 2013

Second revision received January 28, 2014

Third revision received March 3, 2014

Accepted March 28, 2014