IMPROVING THE IMPLEMENTATION AND EFFECTIVENESS OF OUT-OF-SCHOOL-TIME TUTORING

Carolyn J. Heinrich, Patricia Burch, Annalee Good, Rudy Acosta, Huiping Cheng, Marcus Dillender, Christi Kirshbaum, Hiren Nisar, and Mary Stewart

Abstract

School districts are spending millions on tutoring outside regular school day hours for economically and academically disadvantaged students in need of extra academic assistance. Under No Child Left Behind (NCLB), parents of children in persistently low-performing schools were allowed to choose their child's tutoring provider, and together with school districts, they were also primarily responsible for holding providers in the private market accountable for performance. We present results from a multisite, mixed-method longitudinal study of the impact of out-of-school time (OST) tutoring on student reading and mathematics achievement that link provider attributes and policy and program administration variables to tutoring program effectiveness. We find that many students are not getting enough hours of high-quality, differentiated instruction to produce significant gains in their learning, in part because of high hourly rates charged by providers for tutoring. We identify strategies and policy levers that school districts can use to improve OST tutoring policy design and launch improved programs as waivers from NCLB are granted. © 2014 by the Association for Public Policy Analysis and Management.

INTRODUCTION

School districts across the United States are spending millions of Title I dollars on out-of-school time (OST) tutoring for economically and academically disadvantaged students, including large numbers of students with disabilities and English language learners. Under No Child Left Behind (NCLB), public schools that do not make adequate yearly progress (AYP; as defined by their state) for three consecutive years are required to offer children in low-income families the opportunity to receive extra academic assistance (known as supplemental educational services, or SES), consisting of OST tutoring offered primarily by private sector providers.[1] The OST tutoring provisions of NCLB are rooted in the original conceptualization of the ESEA of 1965, which advances the idea of supplemental instruction as a means to improving the quality of instruction for low-income students. Nationwide, 48 percent of schools did not make AYP in the 2010 to 2011 school year, up from 20

[1] As part of the 1994 reauthorization of the Elementary and Secondary Education Act (ESEA), states were required to establish, for the purposes of Title I accountability, a definition of AYP based on student assessment results. Schools are expected to maintain "continuous and substantial improvement" toward those benchmarks, so that all students in the state reach a proficient level of performance against state standards. Each state has defined AYP in its own way, although AYP is the benchmark against which all states identify and assist schools in need of improvement. NCLB's current accountability system has set a goal of 100 percent of students reaching proficiency by 2014.

Journal of Policy Analysis and Management, Vol. 00, No. 00, 1–34 (2014)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.21745

percent in 2006, implying a steady increase in the number of students eligible for OST tutoring under NCLB.[2]

At the start of the 2013 to 2014 school year, 41 states, the District of Columbia, and eight school districts in California had been granted federal waivers that now allow them flexibility to opt out of some core tenets of NCLB (in return for the conditions outlined in their applications). Many of the school districts operating under waivers plan to continue offering OST tutoring, although in modified forms in terms of program design, content, and administration.[3] As school districts exercise newfound authority and flexibility in their efforts to improve OST tutoring services, our research aims to strengthen the research evidence base from which they draw, as well as to support districts in obtaining guidance from research and sharing information on effective practices with their peers in other districts.

We also recognize that other key actors engage with school districts in implementing OST tutoring interventions. Importantly, NCLB explicitly requires parental choice as a lever for improving the quality of OST instruction for economically disadvantaged students (at schools not making AYP), and even following waivers from NCLB, parents (with guidance from districts) are likely to be the primary choosers of tutoring providers to suit their children's individualized needs.[4] NCLB also specifies that the options made available to parents must be "high quality, research-based and specifically designed to increase the academic achievement of eligible children on the State's academic assessments and attain proficiency in meeting the State's academic achievement standards" (U.S. Department of Education, 2005, p. 31). While states typically establish specifications for provider applications and approval, the theory of action for OST tutoring situates the locus of decisionmaking largely at the parent and district levels, presupposing that they have sufficiently accurate and complete information on provider attributes and effectiveness to reap the benefits of choice, as well as adequate capacity or leverage for disciplining the market and rooting out ineffective providers.

Our four-year, multisite mixed methods study of OST tutoring was undertaken to generate rigorous and accessible evidence for informing the design and implementation of "meaningful interventions and support for the lowest-performing schools," an explicit goal of accountability reforms advanced in the anticipated reauthorization of ESEA.[5] Although existing research on OST tutoring interventions has identified some settings and thresholds of tutoring intensity that likely influence tutoring effectiveness, it has generally been limited in its systematic investigation of variables that potentially influence access to and the efficacy of OST tutoring. Our mixed-method, longitudinal investigation in large, urban public school districts probes deeper in examining the attributes of these interventions (as implemented) that

[2] In 24 states, at least half of schools did not make AYP in 2011, with this percentage varying widely by state (from 11 percent to 89 percent). See the Center on Education Policy AYP Results for 2010 to 2011, December 15, 2011. Retrieved on January 17, 2012, from http://www.cep-dc.org/index.cfm?DocumentSubTopicID=48.

[3] We heretofore use OST tutoring as a broader term for tutoring interventions that encompass SES under NCLB.

[4] The guidance states (U.S. Department of Education, 2005, p. 7) the following:

> [A state educational agency] that desires to set program design parameters should ensure that such parameters do not result in the inability of a wide variety of providers, including nonprofits, for profits [local educational agencies], and faith-based and community organizations, from being able to participate as eligible providers, thereby limiting parental choice.

[5] See http://www2.ed.gov/policy/elsec/leg/blueprint/faq/accountability.pdf.

influence their effectiveness, as well as sheds light on how state and district policies and practices can mediate access to and the outcomes of OST tutoring. This is critical information for program development in the post-NCLB waiver environment, where states and districts have newly granted authority to terminate, redesign, and regulate OST programs based on their own priorities and identified student needs.

We begin with a brief review of literature on the efficacy of OST tutoring interventions and then describe our research design, samples, data, and integrated qualitative and quantitative methods. We then present our study findings, which are enriched by the cross-district variation in program and policy implementation that provides important insights into observed relationships between implementation and impacts. We conclude with a discussion of how to improve OST tutoring interventions and the public policies that guide their implementation.

POTENTIAL OF OST TUTORING TO IMPROVE STUDENT ACHIEVEMENT

OST tutoring programs have long been a staple intervention for K–12 students in need of extra academic assistance, and existing studies have explored the relationship of attributes such as program focus, duration, time frame, and student grouping to program outcomes. Lauer et al. (2006) conducted a meta-analysis of 35 peer-reviewed, published studies to estimate effect sizes (e.g., gain scores) of OST tutoring programs. They conclude that OST tutoring can have positive effects on student achievement (in relation to at-risk students who do not participate), and that effect sizes are larger for programs delivering more than 45 hours of tutoring (but smaller for those longest in duration). In a random assignment study of a national after-school program, Dynarski et al. (2004) found no effects on reading test scores or grades for elementary or middle school students, while a follow-up study using these same data (Vandell et al., 2005) reported positive effects on test scores for elementary school students highly active (i.e., participating for 90 or more days) in high-quality programs. A study by Black et al. (2008) of students in grades 2 to 5 randomly assigned to receive either enhanced, adapted models of regular-school-day math and reading instruction in after-school settings or after-school services regularly available at their schools found positive, statistically significant impacts for the enhanced math program on student achievement, but weak evidence of effects on reading achievement, and no effects on student engagement, behavior, or homework completion.

Very few of the earlier studies (the Black et al., 2008 study being an exception) measured program attendance or made the distinction between planned program duration and actual student attendance or engagement. In general, measurement of student contact time or intensity and the quality of instruction in these interventions has been inadequate for understanding program impacts. In addition, OST tutoring programs have faced low and varying attendance rates that are influenced by state, district, and provider policies and supports for registering/enrolling students (Burch et al., 2011; Heinrich, Meyer, & Whitten, 2010; U.S. Department of Education, 2009; Zimmer, Hamilton, & Christina, 2010). The apparent link between student motivation (and other individual and family background characteristics) and engagement in OST tutoring programs poses significant challenges for researchers in identifying the effects of different levels of program intensity or duration and various types and formats of instruction on student achievement.

Looking across nearly a decade of implementation and evaluation of SES under NCLB, few studies find statistically significant, positive effects on student achievement, and where they do, they are generally small (Barnhart, 2011; Burch, 2009;

Deke et al., 2012; Heinrich, Meyer, & Whitten, 2010; Heistad, 2007; Rickles & Barnhart, 2007; Springer, Pepper, & Ghosh-Dastidar, 2009; Zimmer et al., 2007; Zimmer, Hamilton, & Christina, 2010). Estimated effect sizes in these studies for participating students range from approximately 0.05 to 0.09 standard deviations (for reading and math achievement). A recent study by Deke et al. (2012) employed a regression discontinuity design to estimate the average impact of *offering* OST tutoring to eligible applicants who were on the cusp of having access to services in oversubscribed school districts. For students in grades 3 to 8 across six districts, they found no evidence of impacts of offering tutoring to students (near the cut point for an offer) on their achievement in reading or mathematics. They also found no statistically significant impact of *participating* in OST tutoring on student achievement in reading or mathematics. Across their study districts, students received an average of 21.2 hours of OST tutoring over the school year.

Although Deke et al. (2012) concluded that the intensity of services was not significantly related to the estimated size of tutoring impacts in their study, other research (including that of Lauer et al., 2006, discussed above) suggests that reaching some minimum threshold of tutoring hours (i.e., approximately 40 or more hours according to the current evidence base) may be critical to producing measurable effects on students' achievement. Earlier evaluations conducted by Chicago Public Schools (CPS) and Jones (2009) reported larger gains in reading and mathematics for students receiving at least 40 hours of tutoring and for students in grades 4 to 8 who were not English language learners and who received at least 30 hours of OST tutoring. In addition, recent research by Fryer (2012) that examines high dosage tutoring (in an extended school day) for students in low-performing schools—that is, tutoring levels around 200 hours per year or more—finds large effects on student reading and math achievement that are about four to five times the effect sizes typically reported for OST tutoring under NCLB.

In addition to hours of tutoring received, existing research suggests other axes through which increases in academic achievement might be realized. First, a quality OST curriculum is content-rich, differentiated to student needs, and connected to the students' school day (Beckett et al., 2009; Farkas, 2000; Vandell, Reisner, & Pierce, 2007). Second, effective instruction is organized into small grouping patterns (no larger than 10:1 and ideally 3:1 or less or one-on-one), and instructional time is consistent and sustained (Beckett et al., 2009; Elbaum et al., 2000; Farkas & Durham, 2007; Lauer et al., 2006; Little, Wimer, & Weiss, 2008, Lou et al., 1996). Furthermore, instructional strategies are varied (both structured and unstructured, independent and collective), active (not at desk time, worksheets), focused (program components devoted to developing skills), sequenced (using a sequenced set of activities designed to achieve skill development objectives), and explicit (targeting specific skills) (Beckett et al., 2009; Vandell, Reisner, & Pierce, 2007). And beyond elements specific to curriculum and instruction, quality OST programs not only hire and retain tutors with both content and pedagogical knowledge, but also provide instructional staff with continuous support and authentic evaluation (Little, Wimer, & Weiss, 2008; Vandell, Reisner, & Pierce, 2007). Lastly, research suggests the importance of OST tutoring programs actively supporting positive relationships at the classroom level among tutors and students (Durlak & Weissberg, 2007; Vandell, Reisner, & Pierce, 2007), as well as between programs and the surrounding community (Little, Wimer, & Weiss, 2008).

Under NCLB, school districts have not been able to impose requirements on tutors—who do not have to meet *highly qualified* standards or have specific training—and state educational agencies have generally been lax in evaluating providers, setting minimum standards for tutoring quality or requesting essential information on applications for assessing and monitoring quality. Districts and states with waivers now have more leeway to specify program and tutor

requirements and rates per hour charged, and to establish performance-based contracts. However, they need guidance from research in setting these program and performance parameters, particularly given the ongoing challenges of very limited resources for program development and administration and for monitoring and observing providers to understand what is taking place in OST tutoring programs.

RESEARCH DESIGN

The longitudinal, mixed-method design that we employ integrates rigorous, quasi-experimental analysis of OST tutoring program impacts on student achievement with an in-depth, comprehensive examination of the intervention—provider instructional practice in different program models and settings, the nature and quality of tutoring provided, and district-level program administration—in and across four large, urban school districts: CPS, Dallas Independent School District (ISD), Milwaukee Public Schools, and Minneapolis Public Schools. Each of these districts accounts for a disproportionately large share of students eligible or targeted for OST tutoring, and at the start of our study, CPS had one of the largest numbers of students eligible under NCLB, accounting for 10 percent of all recipients in the nation's public schools in 2008 to 2009. Accordingly, student demographics in these school districts reflect those of the larger national (mostly urban) population receiving OST tutoring, that is, high concentrations of economically disadvantaged students, including subgroups with higher levels of academic need/disadvantage (e.g., students with limited English proficiency and disabilities).

Study Samples and Data

We use the targeting criteria determined by the school districts, which have evolved to some extent over time (as do the federal standards for making AYP), to select our study samples each year. These criteria have typically included free lunch-eligible students, English language learners, students with disabilities, and students who are lagging behind their peers academically, as measured by their scores on standardized achievement tests or grade retention. Table 1 provides descriptive statistics on eligible students in these four districts and shows the relative stability in the study population over time (even as district targeting criteria and eligible schools have changed). We draw on five years of data from each study district, including student record, administrative, and test score data from the 2007 to 2008, 2008 to 2009, 2009 to 2010, 2010 to 2011, and 2011 to 2012 school years. Table 2 shows the measures that are commonly available across the districts.

A large number of diverse organizations with widely varying hourly rates, service costs, tutor qualifications, tutoring session length, instructional strategies, and curricula compete for the opportunity to provide OST tutoring. These include national and local organizations, for-profit and nonprofit providers, online and off-line providers, those offering services on-site at schools (and off-site), and as in CPS, some school districts engaging directly in the provision of OST tutoring. NCLB explicitly discouraged state and local educational agencies from taking any actions that might limit the supply of providers or range of choices available to parents, and they likewise could not specify or constrain hourly rates charged by providers. Our study sample includes close to 200 unique providers of OST tutoring, as well as some that have offered services in more than one (or all) of our study districts. Data on provider characteristics and program features (from state and districts sources

Table 1. Characteristics of students eligible for out-of-school tutoring in study districts.

All eligible students	Chicago public schools				Dallas independent school district			
	2008 to 2009	2009 to 2010	2010 to 2011	2011 to 2012	2008 to 2009	2009 to 2010	2010 to 2011	2011 to 2012
Number of students and characteristics	88,353	87,542	101,930	245,616	35,612	30,774	35,026	39,091
Asian (%)	1	2	2	2	1	1	1	1
Black (%)	53	49	42	43	34	31	30	31
Hispanic (%)	44	47	53	51	62	64	65	62
White (%)	2	2	2	3	3	4	4	3
Other race (%)	0	0	1	1	0	0	0	1
Female (%)	49	49	49	50	48	48	48	48
ELL (%)	12	12	16	18	21	19	16	20
Free lunch (%)	100	100	100	99	67	79	74	60
W/disabilities (%)	14	13	12	13	12	12	11	11
Attended SES last year (%)	26	42	8	13	16	15	37	28
Absent last year (%)	6	4	5	5	7	9	7	6
Retained this year (%)	4	2	2	2	0	7	8	8
Number of students and characteristics	Milwaukee public schools				Minneapolis public schools			
	11,992	26,798	16,439	20,905	10,963	15,769	16,444	15,906
Asian (%)	5	4	4	4	11	9	9	9
Black (%)	68	69	68	68	48	47	46	45
Hispanic (%)	17	20	20	20	28	29	28	26
White (%)	8	5	8	6	6	8	9	12
Other race (%)	3	3	0	0	6	7	7	8
Female (%)	48	47	46	46	51	50	50	49
ELL (%)	6	10	12	10	34	36	33	36
Free lunch (%)	83	87	88	90	99	100	100	98
W/disabilities (%)	21	22	22	24	17	18	18	18
Attended SES last year (%)	11	6	14	8	13	7	16	16
Absent last year (%)	16	15	16	13	8	8	7	4
Retained this year (%)	13	11	12	12	2	6	2	5

and our own data collection) are linked to recorded/invoiced hours of tutoring for each student and other student-level data for analysis.

The qualitative data that we collected in this study over the 2009–2010 to 2011–2012 school years include (1) observations of full tutoring sessions ($n = 123$) using a classroom observation instrument designed to capture key features of instructional settings; (2) interviews with provider administrators ($n = 55$) about the structure of instructional programs, choice of curricula and assessments, challenges in implementation, and choices in staffing; (3) interviews with tutoring staff ($n = 69$) about instructional formats, curriculum, adaptations for special student needs, and staff professional background and training; (4) interviews with district and state administrators ($n = 31$) involved in program implementation; and (5) parent focus groups ($n = 155$) with parents of students who were eligible to receive OST tutoring. The documents analyzed include formal curriculum materials from providers; diagnostic, formative, or final assessments used; and policy documents on federal, state, or

Table 2. Description of variables available for empirical analysis across the study sites.

Core control variables	Description	Site-specific details
Core controls	Eligible for OST tutoring Registered for OST tutoring with district Received tutoring (nonzero hours)	
Student identification and enrollment information	Student ID District code District assigned local identification number	
Student demographic information	Student gender Student ethnicity (black, white, Hispanic, Asian, or other) Student age Limited English proficiency/English language learner (ELL) indicator Economic disadvantage status (indicated by free or reduced lunch) Enrolled in special education program during school year Retained in the same grade as the prior school year Period for which attendance is recorded	
Attendance and absence information	Percent of days absent in prior school year Number of days absent during the reporting period	
Basic treatment measures	**Description**	**Site-specific details**
Treatment information	Attended any tutoring in prior school year Hours of tutoring received (invoiced) Tutoring provider	
Primary outcome measures	**Description**	**Site-specific details**
Reading measures	Change in reading scores (standardized with district average test scores) Change in reading scores (standardized with average of SES eligible test scores)	Dallas: TAKS, STAAR, Chicago: ISAT, ITBS Milwaukee: WKCE Minneapolis: MCA-II
Math measures	Change in math scores (standardized with district test score averages) Change in math scores (standardized with average of SES eligible test scores)	Dallas: TAKS, STARR Chicago: ISAT, ITBS Milwaukee: WKCE Minneapolis: MCA-II, MTELL

district policies concerning the implementation of SES. Appendix A describes these data sources in greater detail.[6]

[6] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

By design, our research has sought to tightly integrate the application of qualitative and quantitative methods in data collection and analysis in order to better understand the mechanisms or pathways to tutoring program impacts. For example, we have used mixed methods to optimize our sample, with quantitative data facilitating our identification of parameters (e.g., student market share, cross-site enrollment) that guide the selection of tutoring providers subsequently observed in the field research. We have also improved the sensitivity and appropriateness of our core instruments and measures through qualitative–quantitative integration. To measure treatment and student participation in tutoring, we rely on both large-sample, standardized measures (i.e., invoiced hours of OST tutoring) in the quantitative analysis, and observations of tutoring practice, interviews, and analysis of curriculum in the qualitative work to understand what is happening in an hour of tutoring in practice, that is, what comprises an invoiced hour in terms of instruction. We have learned that underlying the invoiced hour is a much more complicated story of instructional time that includes incomplete record keeping, students leaving early or arriving late, tutoring time spent on noninstructional activities, technical/materials difficulties, and other issues. The interviews and observation data have also revealed important differences within digital[7] tutoring formats. This information is critical in refining both our measures and interpretations in data analysis, and accordingly, in increasing the validity of our research.

Qualitative Research Methods and Analysis

The qualitative research is grounded in two key principles: (1) a sustained focus on instructional setting, where we map backward from program characteristics to classroom and school-level characteristics that contribute to program accessibility, quality, and impact (in particular, teacher–tutor, school–leader–teacher interactions), as well as district, state, and federal policy and program characteristics that are linked to these factors and mediate impacts; and (2) a sharp focus on the system of OST implementers, that is, classroom teachers, providers, parents, tutors, school personnel, and district and state staff, that enable or impede the effectiveness of OST tutoring interventions. In the context of ongoing educational reform efforts, we seek to understand the critical exchanges of information and resources between and across these stakeholders around the implementation of OST programming, and to illuminate perspectives of these multiple stakeholders as they shape and are shaped by program implementation.

A centerpiece of our mixed-method, qualitative work is a standardized observation instrument we developed to more precisely capture the nature of supplemental instruction.[8] Systematic analysis of structured observation protocols offers critical insight into the evaluation of accountability-based programs (Pianta & Hamre, 2009). We link observation data to tutor interview data and content analysis of planned and enacted curriculum in the same instructional setting. The instrument has the capability of not only providing descriptive information on instructional materials and teaching methods *in use*, but also detecting the effects of different kinds of formats, resources (curriculum materials, staffing), and instructional methods on students' observed levels of engagement. The observation instrument includes indicator ratings at two, 10 to 15 minute observation points, as well as a rich

[7] We define a *digital provider* as one that uses a digital platform (i.e., software or live tutor via a computer or hand-held device) as an intentional, integral part of its instructional strategy.
[8] A copy of the observation instrument is available at http://www.sesiq2.wceruw.org.

description in the form of a vignette and follow-up information provided by the tutor(s).[9]

The observation data are subsequently categorized into clusters of indicators, organized by areas of OST best practice: varied, active, focused, targeted, relationships, tutor knowledge, differentiation, and student engagement. This clustering of qualitative indicators allows us to see which best practices are predominant in observations and which are rare or missing. Although the observation instrument ratings use a number rating system, the process is fully qualitative in terms of coding and clustering the indicators under each best practice area, as we look for patterns and outliers in the data that are then used in refining the instrument. OST cluster numbers are calculated by adding the total ratings for each indicator in each cluster and dividing that sum by the total possible ratings. We triangulate these data with narrative vignettes, or rich descriptions, of the instructional setting to give context to the number ratings.

We use a constant comparative method (both within and across methods) to develop and refine our understanding of patterns and dissimilarities in tutoring practices across providers. The same data are analyzed and discussed simultaneously by different researchers in an effort to consider and develop multiple interpretations of events observed. We also create opportunities through the research cycle to share early observations with key stakeholders to make sure we are capturing local realties and distinctions. Throughout the process, we seek to examine possible trends in instructional settings that may help in understanding the local challenges in policy implementation and strategies to address them. Analytic codes developed from these patterns and in response to the research questions are then reapplied to interview, observation, and archival data to establish findings. Data analysis occurs both concurrent to and after data collection.

Quantitative Research Methods and Analysis

In evaluating OST tutoring impacts, we are faced with the classic evaluation problem that it is necessary to identify both actual participant outcomes and the outcomes that would have occurred for them absent participation. We define Y_1 as the test score for a student following participation in OST tutoring, and Y_0 as the test score for that student over the same period in the absence of participation. It is impossible to observe both measures for a single student. If we specify $D = 1$ for those who participate in OST tutoring and $D = 0$ for eligible students who do not participate, the outcome we observe for an individual is

$$Y = (1 - D) + D. \tag{1}$$

Evaluations employing random assignment methods ensure that the treatment is independent of Y_0 and Y_1 and the factors influencing them. Random assignment was not an option in this study, however, given the federal mandate to make OST tutoring available to as many eligible students as funding allowed and relatively low early program take-up rates.

Where D is not independent of factors influencing Y_0, participants may differ from eligible nonparticipants in many ways besides program participation, so the simple difference in outcomes between participants and eligible nonparticipants

[9] We conduct regular reliability trainings with the qualitative research team to ensure consistency in ratings. Validity of the instrument is ensured by the development process, whereas its structure and content is based on well-tested, existing observation instruments for OST, existing literature on the best practices for OST, and the theory of action in the supplemental educational services policy.

will not necessarily identify program impacts. If we assume that given measured characteristics (a set of conditioning variables, *X)*, participation is independent of the outcome that would occur in the absence of participation, the effect of OST tutoring on participants conditional on X can be written as

$$E(Y_1 - Y_0|D = 1, X) = E(\Delta Y|D = 1, X) = E(Y_1|D = 1, X) - E(Y_0|D = 0, X) \quad (2)$$

where $Y_1 - Y_0 = \Delta Y$ is estimated to be the program effect for a given student, and the expectation is across all participants with given characteristics. (This is the conditional independence assumption, or the assumption of unconfoundedness.) Regression adjustment and matching methods are all based on some version of equation (2), but they differ in the methods used to obtain estimates of $E(Y_1|D = 1, X)$ and $E(Y_0|D = 0, X)$.

The primary strategy for quasi-experimental estimation of average OST tutoring impacts that we employ is value-added modeling with school fixed effects. In these models, our comparison groups consist of students eligible for OST tutoring (in each district) who do not receive tutoring. We have found a high degree of consistency in estimates produced by alternative value-added and fixed effects model specifications (Heinrich & Nisar, 2013), as further discussed in Appendix B; therefore, we focus our discussion later on the results of the value added models with school fixed effects.[10] In addition, we also estimate generalized propensity score (GPS) matching models to assess the effects of different dosages (hours) of tutoring. In these models, we include only students who received at least an hour of tutoring and model their selection into alternative dosages of OST tutoring.

The outcome measures in these models are the achievement gains made by students (from one school year to the next) in mathematics and reading, measured as the difference in their (scale) scores on standardized tests administered both before and after their participation in OST tutoring. In addition, we calculate effect sizes by standardizing these student gain scores relative to the district average gains in student math and reading achievement. In our estimation, we have also tested alternative approaches to controlling for students' prior academic performance. For example, a common strategy is to include the pretest score (Y_{t-1}) on the right hand side of the model (with other conditioning variables, X) as a predictor of student achievement (measured by the test score after tutoring, Y_1). We discuss these alternative specifications and results from their estimation as well in Appendix B.[10] Here, we briefly describe the basics of our primary estimation strategies and their assumptions below.

Value-Added Model with School Fixed Effects

The value-added model we employ allows us to control for other classroom and school interventions that are fixed over time. For example, if there is a reading intervention at a school and those students also receive tutoring, failing to control for the intervention (school fixed effect, π_s) would bias the results. We estimate

$$A_{jst} - A_{jst-1} = a\text{OST}_{jt} + \beta X_{jt-1} + \pi_s + \mu_{gt} + E_{jst} \quad (3)$$

[10] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787. All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

where A_{jst} is the achievement of student j attending school s in year t; OST_{jt} is an indicator function if the student, j, attended tutoring in year t; X_{jt-1} are student characteristics that include student demographics, percent absent in prior year, retained in prior year, and attended tutoring in the prior year; π_s is school fixed effect; μ_{gt} are grade by year fixed effects; and E_{jst} is the random error term. Identification in this specification comes from the average gain in student achievement after controlling for student characteristics and school and grade year effects. The outcome measure is the achievement gain made by students, which accounts for the possibility that students with similar characteristics might enter OST tutoring with different underlying achievement trajectories (as reflected in their prior test scores).

Matching Methods

Matching methods are designed to ensure that estimates of program impacts are based on outcome differences between comparable individuals. In the sample of participants and eligible nonparticipants, $P(X)$ is the probability that an individual with characteristics X is a participant. Rosenbaum and Rubin (1983) showed that $Y_0 \perp\!\!\!\perp D|\text{X} \Rightarrow Y_0 \perp\!\!\!\perp D|P(X)$, which implies that for participants and eligible nonparticipants with the same $P(X)$, the distribution of X across these groups will be the same. That is, we assume *conditional independence:* There is a set of observable covariates, X, such that after controlling for these covariates, potential outcomes are independent of the treatment status.

If students receive varying dosages of tutoring, as we observe, then the average treatment effect estimated by conventional estimators will not capture heterogeneity in effects that may arise. In light of this, and with sufficient data (distributed normally) on OST tutoring dosages, we estimate GPS models of program impacts, in which participants are matched with individuals in a comparison group based on an estimate of the probability that the individual receives a given dosage of treatment (the GPS). The GPS approach assumes that selection into levels of treatment (tutoring) is random, conditional on a set of rich observable characteristics; that is, the level of participation is independent of the outcome that would occur in absence of participation. If the model assumptions are satisfied, it is possible to use GPS to estimate the average treatment effects of receiving different dosages of OST tutoring, thereby allowing for the construction of a *dose–response function* that shows how treatment exposure relates to outcomes.

We follow Hirano and Imbens (2004) and define $\mathcal{T}$ as the set of all treatment levels (hours of OST tutoring attended); T as a specific treatment (hours) level, and the treatment interval as $[t_0, t_1]$, so that $T \in [\text{t}_0, \text{t}_1]$). We calculate the average dose–response function, $\mu(t) = E[Y(t)]$, assuming unconfoundedness; that is, after controlling for X, mean outcomes for comparison cases are identical to outcomes of participants receiving T hours of tutoring. The GPS, R, is defined as $R = \text{r}(T, X)$, so that under this assumption and within strata with the same value of $\text{r}(T, X)$, the probability that $T = t$ does not depend on the value of X (Hirano & Imbens 2004, p. 2). We estimate values of the GPS using maximum likelihood, assuming the treatment variable is normally distributed, conditional on the covariates X: $g(T) \mid X \sim N\{h(\gamma, X), \sigma^2\}$: $\hat{R}_i = [2\pi\sigma^2]^{(-0.5)} \exp[(-(2\sigma^2)^{-1})[g(T_i) - h(\gamma, X)]]$. The balancing properties are checked and the conditional expectation of Y (the response), given T and R, is estimated. (See Appendix B for further details on the GPS estimation).[11]

[11] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

Table 3. Average impacts of OST tutoring by school district, year, and student group on reading and math achievement (gains).

	Reading		Math		Reading		Math	
All students	No. of students with gain scores	Effect size	No. of students with gain scores	Effect size	No. of students with gain scores	Effect size	No. of students with gain scores	Effect size
	2008 to 2009 school year				2009 to 2010 school year			
Chicago	61,171	**0.043**	61,464	**0.046**	63,506	**0.094**	63,773	**0.053**
Minneapolis	2,862	−0.202	1,400	−0.011	1,602	−0.202	789	−0.011
Milwaukee	4,697	−0.079	4,772	−0.048	1,841	−0.079	1,870	−0.048
Dallas	9,294	−0.109	9,294	−0.076	14,106	**0.111**	13,807	**0.127**
	2010 to 2011 school year				2011 to 2012 school year			
Chicago	205,187	**0.075**	204,094	**0.064**	68,541	**0.042**	68,411	**0.045**
Minneapolis	5,025	**0.144**	5,045	**0.191**	4,247	−0.037	4,298	0.050
Milwaukee	2,826	0.021	2,831	−0.043	3,668	−0.020	3,663	0.031
Dallas	13,428	0.016	13,333	0.016	14,670	0.011	14,361	0.054

Note: Statistically significant impacts on student achievement (at $\alpha = 0.05$) are reported in bold.

STUDY FINDINGS

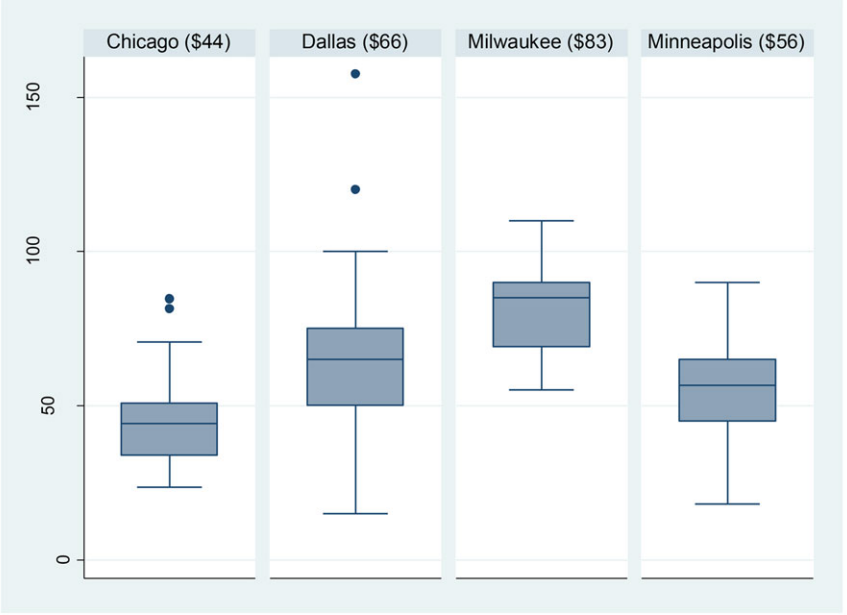
Average Impacts of OST Tutoring

We first present findings on the average impacts of OST tutoring across four years in the four study school districts (from the value-added models with school fixed effects) for all students receiving tutoring (see Table 3). The reported coefficients are effect sizes—that is, the change, measured in standard deviations from district average reading and math test scores, in an average student's outcome (gain) that can be expected if the student participates in OST tutoring. Statistically significant coefficient estimates (at $\alpha = 0.05$) are reported in bold.

The results in Table 3 indicate that only in CPS do we *consistently* find statistically significant average impacts of OST tutoring on students' reading and math achievement. We observe just a few statistically significant effects of tutoring in the other districts—in the 2009 to 2010 school year in Dallas ISD and in Minneapolis Public Schools in 2010 to 2011. The magnitude of effect sizes (where observed)—approximately 0.04 to 0.12 (with the exception of larger effect sizes in Minneapolis Public Schools in 2010 to 2011)—are comparable to those estimated in other studies that have identified impacts of OST tutoring under NCLB (Heinrich, Meyer, & Whitten, 2010; Springer, Pepper, & Ghosh-Dastidar, 2009; Zimmer et al., 2007). Although for brevity, we do not separately present the results for the subgroups of English language learners and students with disabilities, we again only observe consistent positive impacts of OST tutoring for these subgroups in CPS, and for students with disabilities, the effect sizes are always smaller (approximately half the size at about 0.03 SD).[12]

Figure 1 shows the hourly rates charged by tutoring providers in the study districts in the 2011 to 2012 school year, as well as the average number of hours of tutoring

[12] These results are available from the authors upon request.



Note: The average hours of tutoring provided by each district, shown below, correspond to the average hourly rates charged for tutoring by each district above. Students in the districts with lower average hourly tutoring rates receive higher numbers of hours of tutoring, on average.

Average hours of OST tutoring received by participating students by school district and year				
District	**2008–09**	**2009–10**	**2010–11**	**2011–12**
Chicago	38.8	38.7	39.0	35.7
Dallas	21.8	35.2	17.9	14.8
Milwaukee	25.9	28.2	28.3	21.7
Minneapolis	26.7	28.7	31.9	27.2

Figure 1. Average Provider Hourly Rates in 2011 to 2012 School Year and Hours of OST Tutoring Provided by District, 2008 to 2012.

received by students by district and school year (from 2008 to 2009 through 2011 to 2012). In CPS, where students have routinely reached thresholds of 36 to 39 hours of OST tutoring (on average) and we consistently observe positive impacts of OST tutoring, provider hourly rates are the lowest on average (at $44 per hour). In practice, the number of hours students attend is directly influenced by the rate per hour charged by tutoring providers and the dollars allocated per student by districts.[13] For example, one district in our study allocated approximately $1,300 per student for tutoring; as over 70 percent of these students received tutoring from a provider charging $75 or more per hour, the maximum hours of OST tutoring a student could receive was about 18 hours over the school year. We have observed

[13] Districts set aside up to 20 percent of their Title I funds (under NCLB) for tutoring, and the total amount they can allocate per student depends on the number of eligible students prioritized for services.

provider hourly rates as low as $13.25 and as high as more than $157 per hour in our study districts, and district funding allocations for OST tutoring have ranged from approximately $1,100 to $2,000 per student.[14]

The positive impacts of OST tutoring that we observe in Dallas ISD in 2009 to 2010 and Minneapolis Public Schools in 2010 to 2011 both coincide with *natural policy experiments*, in which limited-time policy or program changes directly increased the number of hours of OST tutoring students that students received only in those sites and years. In Dallas ISD, the district intentionally used federal stimulus funds in 2009 to 2010 to increase the allotted district expenditure per student and thereby boost the number of hours of tutoring students received. Figure 1 shows that average hours of OST tutoring increased from approximately 22 hours in 2008 to 2009 to 35 hours in 2009 to 2010, and then fell by half and more in the subsequent two school years. In Minneapolis, the district introduced a new program in 2010 to 2011 for a subset (approximately one-sixth) of OST tutoring participants that, as informed by the current evidence base, compelled providers to deliver more (at least 40) hours of tutoring. Students in this trial program received about 20 more hours of tutoring, which increased the overall average number of hours tutored (only for that year) to about 32 hours. In sum, a common pattern in these results (in Table 3 and Figure 1) is that we only observe positive program impacts in districts and years where average hours of OST tutoring exceeded 30 hours.

These findings suggest of a strong relationship between hours of OST tutoring and program effectiveness. With other analytical techniques, we can more precisely examine the relationship between hours of OST tutoring and effects on student achievement, including the possibility that there are growing or diminishing returns to tutoring as the number of hours of tutoring increases. We further explore the linkages between hours (or *dosages*) of OST tutoring and program impacts using GPS matching.

Effects of Varying Dosages of Tutoring

We estimated GPS models to investigate the relationship between hours of OST tutoring and students' reading and math gains for each district in our study, but only in CPS did we have sufficient numbers of students reaching higher levels of tutoring to estimate effects precisely and achieve covariate balance at a range of different tutoring dosages. Therefore, we focus here on the results of the GPS estimation for CPS, using hours of tutoring accumulated by students over three school years. Where there were an adequate number of observations in other study sites, primarily in Minneapolis Public Schools, we see similar patterns of effects across varying dosages of OST tutoring, albeit with wider confidence intervals due to the smaller number of observations at each level (results available from the authors).

Appendix B provides more details and programming code for the estimation of the GPS models.[15] The models specify treatment levels (dosages) from 10 to 80 hours of tutoring, with the dose–response estimated at five-hour increments in tutoring (i.e., estimated effects at 10, 15, 20, 25, etc., hours of tutoring, up to 80 hours). Although covariates (controlling for student demographics, school attendance, past performance, grade retention, and grade level) did not fully balance at every cut

[14] Farkas and Durham (2007) likewise found that high rates charged by providers limited the number of hours of tutoring students received and tutoring program effectiveness.

[15] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's Web site and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787.

point in the distribution[16]—particularly at the highest and lowest levels of OST tutoring where observed dosages were more sparse—the estimated (linear) dose–response functions show a clear relationship between hours tutored and increases in student achievement, and the effects are precisely estimated.

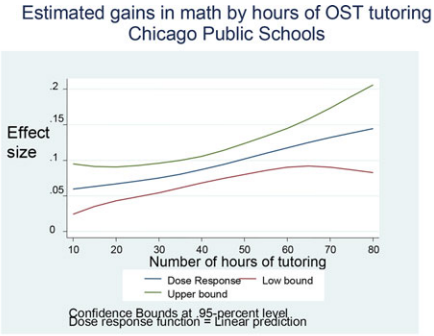
The GPS results are presented in both graphic and tabular form in Figure 2, with hours of tutoring measured along the horizontal axis in the graphs and the estimated effect sizes for a given level of tutoring shown on the vertical axis, as well as in the table of dose–response estimates and standard errors. The middle (solid) line in the graphs shows how effect sizes (i.e., average gains in reading or math, measured in standard deviations from district averages) change as tutoring dosages increase. The dashed lines are the confidence intervals (upper and lower bounds for the effects). For student achievement in math, gains from OST tutoring continue to grow with increasing tutoring dosages through 80 hours (with effect size gains of 0.005 to 0.006 for each five-hour increment of tutoring). All estimates are statistically significant, although they are most precise for the 40 to 50 hour range, where the number of observations is densest. For student achievement in reading, the gains to additional hours of tutoring appear to level off at the 60-hour threshold and then decline with additional hours of tutoring (see the marker on the solid line in the graph). These results indicate that if most students got as many as 60 hours of OST tutoring, reading gains from tutoring would increase (close to doubling from effect sizes at 25 hours), and math gains would be even larger for additional hours of tutoring beyond 60 hours. In other words, the GPS analysis confirms that larger student achievement gains could be realized if more students received higher dosages of OST tutoring.

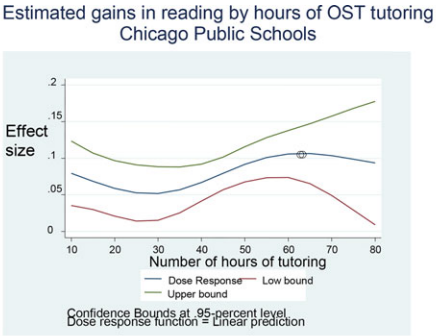
What Takes Place in an Invoiced Hour of OST Tutoring?

In addition to the quantity of OST tutoring received by students, our in-depth qualitative study probes the quality of OST tutoring, or what is happening in practice in an invoiced hour of tutoring, to understand how OST tutoring effectiveness could be increased through program improvements. More specifically, examining key elements of tutoring program models as implemented and assessing their fidelity to evidence-based best practices, we identify how policy and implementation mediate program impacts, and how this varies across districts and different provider settings, formats, and approaches to tutoring. Our results are based on 123 observations of OST tutoring sessions across a range of providers that include digital, in-home, in-school, and community-based tutors; for-profit, not-for-profit, district-provided, and faith-based organizations; providers with large market share (in terms of students served) and with higher than average levels of student attendance; and providers advertising services to students with disabilities and English language learner populations.

In general, the model of OST tutoring commonly observed took the form of traditional academic learning environments, with students being tutored in tested subjects—mathematics and reading—and typically instructed in a whole group format with more than one student and one focal activity. Students receiving tutoring who might learn best via project-based learning, arts integration, or links to community-based activities encountered few opportunities of this sort across the study districts. Furthermore, very few tutors with training or experience in working with English language learners or students with disabilities were present during

[16] Balance was not achieved for some demographic characteristics such as English language learner, other race, percent absent, and grade level in cut points of the distribution with the fewest cases.





Note: The middle line in the above graphs shows how effect sizes (i.e., average gains in reading or math) change with each additional hour of tutoring. The outer lines are confidence intervals (or bounds for the effects). Below are the estimates of the effect sizes for tutoring dosages (in five-hour increments), which make up the plotted relationships in the figures above.

GPS dose-response estimates of OST tutoring dosage effects				
	Math gains		**Reading gains**	
Dosage of OST tutoring in hours	Effect size estimate	Standard error	Effect size estimate	Standard error
10	0.063	0.014	0.078	0.016
15	0.068	0.012	0.072	0.013
20	0.074	0.011	0.067	0.012
25	0.079	0.011	0.067	0.012
30	0.084	0.011	0.072	0.012
35	0.089	0.010	0.082	0.010
40	0.094	0.009	0.096	0.009
45	0.098	0.008	0.111	0.010
50	0.103	0.009	0.122	0.011
55	0.109	0.010	0.129	0.012
60	0.114	0.012	0.129	0.013
65	0.120	0.014	0.125	0.015
70	0.126	0.016	0.117	0.019
75	0.132	0.020	0.109	0.024
80	0.138	0.023	0.102	0.030

Figure 2. Generalized Propensity Score (GPS) Matching Results Showing Relationship Between Gains in Student Reading and Math Achievement and Tutoring Dosage (in Hours).

tutoring, and curriculum nor instruction were rarely tailored in any way to the unique needs of these students. See Burch and Good (in press) for a more detailed description of the nature of the OST instructional landscape (i.e., SES) under NCLB.

Advertised Versus Instructional Time

We frequently observed differences between the advertised time of tutoring sessions and the actual instructional time. Providers are required to advertise the average length of their sessions, and districts are invoiced at an hourly rate based on the

Table 4. Average advertised versus instructional time (in minutes by format; median in parentheses) across all districts from 2009 to 2012.

Format	Advertised time	Instructional time	Difference
Digital ($N = 26$)	83.25 (60)	64.23 (60)	11.95 (9)[a]
Home ($N = 21$)	62.86 (60)	59.43 (58)	3.43 (3)
School ($N = 79$)	99.10 (110)	80.14 (81.5)	18.96 (14)
Community ($N = 20$)	123.16 (120)	90.11 (75)	29.10 (22.5)[b]

[a]The discrepancies between the calculated average difference between advertised and instructional time and the difference between average times is due to the fact that software-based digital program duration is controlled by the student alone, thus not providing an advertised time for those sessions. Calculated average differences between times only take into account sessions that have both an advertised and an instructional time, but the instructional times listed here include sessions that do not have advertised times.

[b]The discrepancies between the calculated average difference between advertised and instructional time and the difference between average times is due to two issues: a few sessions did not have advertised times, and one of the observed sessions did not have a precise observed instructional time. These values were not used in the calculations for average difference, but the instructional times in the first group of sessions and the advertised time in the second example were included in the calculations for average advertised and average instructional times.

time students spend in tutoring. In our sample, advertised sessions ranged from 60 to 240 minutes. Irrespective of the format, students tended to receive less instructional time than what was advertised by providers, although the magnitude of these differences varied by format. As displayed in Table 4, tutoring completed in the student's home most closely matched instructional time with advertised time (approximately three minutes difference on average). In school and community settings, average instructional time was often considerably less than average advertised time: approximately 19 minutes in the case of in-person, school-based tutoring and approximately 29 minutes in the case of in-person, community-based tutoring. Digital tutoring averaged a difference of 12 minutes.

Our fieldwork also offers insight into possible reasons for these discrepancies. In school-based OST tutoring, the format necessitates administrative tasks (e.g., rosters, snacks, transportation). One tutor in a school-based program remarked, "By the time you go pick up the students and bring 'em to your room, they lost about five minutes. You know? Then you pass out the materials. I probably have 'em for about 55 minutes." In addition, tutoring sessions compete with other activities (such as sports teams) for time. On average, there tend to be larger numbers of students, and time is needed for these students to transition from school dismissal to the tutoring sessions. In some community settings, logistics of transportation (e.g., handing out bus tokens, making sure that students get outside to meet the bus, or checking in with families as the provider picked up and dropped off students) sometimes prevented sessions from lasting for the full, advertised time. School and community settings also often include food, which requires extra time and is not the case in digital or in-home sessions. Regardless of the reason, in sessions where there are demands on tutors to conduct activities other than instruction, participating students are likely not getting the full instructional treatment advertised.

Attendance Flux

In 38 percent of observations with two or more students—primarily nondigital, school-based, or community-based settings—students that started a session were observed missing part of the session or leaving the tutoring session altogether, or

students came in late. We call this "attendance flux." Observation data indicated a large number of tutoring sessions had considerable student attendance flux, as measured by comparing the number of students observed in Observation Point A with the number of students observed in Observation Point B. When these numbers were not the same, we counted this observation as having attendance flux. Of the 84 observations with two or more students, 32 (38 percent) had (attendance) flux. Four of the 32 sessions with flux took place in community-based settings (four of 12 total community-based observations with two or more students), and 24 of the 32 sessions with flux took place in school-based settings (24 of 61 total school-based observations with two or more students). One of four digital sessions had flux, and zero out of two home-based sessions with two or more students had flux.[17]

As noted above, the higher proportion of school-based attendance flux may reflect competition with other school-based activities. Through observations as well as interviews with both tutors and provider administrators, we know that school-based tutoring programs often compete with other after-school programs (e.g., athletics and clubs) for students' time. For example, in one school-based tutoring observation, we noted a handful of students leaving a tutoring session early to attend a school-sponsored club that meets weekly to improve students' self-esteem. In addition to decreased instructional time during sessions, students who move frequently in and out of sessions may realize fewer benefits of tutoring.

Variation Within OST Tutoring Providers

We also observed considerable variation in the treatment or instructional program *within* provider. The theory of action behind OST tutoring under NCLB is that variation *between* providers creates a competitive marketplace from which parents can choose the most appropriate program for their students' needs. Variation within providers confounds the assumption that the axis of parental choice lies on the provider level and also may complicate efforts to evaluate tutoring program effects at the provider level.

For example, sessions of very different instructional styles and quality were observed for one provider who offers services both in schools and homes. In one session at a school site, the tutor worked with three students together for one hour on a variety of math activities all focused on the same concepts around long division. This tutor was also the math specialist for the school and incorporated a number of activities and strategies from her day school resources to engage students in active learning. On the other hand, a tutor from the same provider worked with one student at home for two hours. She was not a certified teacher, although had coursework and experience in tutoring. She relied exclusively on the printed worksheets from the provider and jumped from concept to concept, even from math to reading, depending on the worksheet. The student was not actively engaged.

As this example illustrates, there is intraprovider variation in both instruction and in curriculum materials, as they come from a variety of formal (Web site or materials directly from provider administrators) and informal sources (tutors own resources or students' work from day school). The in-use curriculum often included formal materials supplemented by materials from the tutor, the latter being occasionally inconsistent with the formal curriculum. On a more encouraging note, tutors were observed engaging with students in a predominantly positive way across

[17] Seven of the 84 total observations are categorized as both school-based and digital formats, as there were students using either digital or in-person tutoring components in the same classroom. These seven observations were therefore removed from the analysis of attendance flux or instructional time by format, but remain as part of totals.

districts and formats, which is a critical part of meeting the broad social and emotional needs of students. Specifically, tutoring sessions rated highly on indicators of best practices such as provide constructive criticism, encourage participation from disengaged students, and listen actively and attentively to students. In addition, tutoring consistently occurred in small groups, approximately 80 percent of all sessions had a student–tutor ratio of less than 4:1.

OST Tutoring for Students with Special Needs

In light of the quantitative findings of fewer and smaller effects of OST tutoring for students with special needs (in this case, including both English language learners and students with disabilities), we looked more closely at the nature of the intervention in practice (from identification and registration to assessment and instruction) for these two subgroups of students. One of the central issues concerning students with special needs under NCLB is confusion over who is legally responsible for serving English language learners and students with disabilities. OST tutoring providers depend on parents, teachers, and districts to share student assessment data in order to know what types of students are coming their way and to have staff prepared to meet any challenges and tailor services for students with disabilities or English language learners. Across our study districts, we encountered conflicting evidence regarding how providers are informed of students' English language learner or disability status. Some districts give student Individualized Education Programs (IEPs) to providers if the provider requests additional information regarding a student, while other districts provide student IEPs through the district/provider database, which may conflict with student's confidentiality as governed by the Individuals with Disabilities Education Act (IDEA) or the Family Education Rights and Privacy Act (FERPA). Considering that parents voluntarily enroll their students, it might be legally acceptable for a district to include a parental consent section regarding educational record disclosure to providers on the application, as a district in our study currently does. However, it is not clear whether this level of parental consent is sufficient to meet FERPA requirements.[18]

Data sharing and communication among providers and school/district personnel is in many cases dependent on the relationship between the provider and teachers at the school level. If school-day teachers are employed by the provider as tutors, they can more easily negotiate access to IEPs and personally network tutors and teachers to discuss students' specific needs. If there is no existing relationship between the school and the provider, communication between school personnel and tutors is more difficult to facilitate. Furthermore, some schools and districts have strained relationships with providers, where providers feel schools are not welcoming and supportive of the SES program, and schools feel put upon to implement another intervention with no additional funds with which to manage it. This can prevent providers from receiving up-to-date information on participating students' educational needs. The result is that in most OST tutoring sessions, tutors have little or no knowledge of their students' specific instructional needs.

Providers and their tutors might get information about students' special needs directly from parents via phone calls, e-mails, or in-person meetings. One provider administrator explains this in detail:

[18] Additional regulations from IDEA regarding confidentiality include 34 CFR § 300.123; 34 CFR § 300.622; and 34 CFR § 300.623.

I: Let's take the example of a student with a disability, at what point do you know about that disability.

R: We usually don't know unless the parent calls ahead of time, like says, hey, you know, I have some questions about your program. If they call me, I can collect some information. If I see them at a school event, I tend to ask those questions, but we aren't supposed to collect information at school events. I can write down a name like Jose, and that he has an IEP, he's at [school name]. Okay, well it's Jose, say Jose registers at [school name], then I have a pretty good chance of knowing that that's the Jose that has the trouble with math and has the IEP. But unless I interact with the parent at the school, or, you know, one of my employees have, or that they've called me and I've been able to retrieve that information ahead of time, mostly we do not know.

However, while most parents are supportive and want their student to receive tutoring, some providers do not offer such services as in-person home visits or translation for parent phone calls that may be necessary for some families. As the provider administrator quoted above mentioned, some districts offer provider fairs to connect parents with provider representatives and tutors; however, again, these representatives may not have the capacity to communicate with parents in their native language. Many districts try to identify the variety of languages spoken in their district and find translators for each one; however, providers in all districts—and particularly multidistrict providers—have noted the difficulty of keeping translators on staff, having translators for all applicable languages, and finding tutoring staff who are both bilingual and have special education training.

In addition, some OST tutors with special education or English as a second language instructional backgrounds have been trained in appropriate diagnostics to identify students' needs, but these teachers do not always get matched with students needing their particular areas of training and experience. A primary reason for this is that tutoring providers may not have access to school records or staff with knowledge about students' needs, and therefore cannot match students and tutors accordingly. As a result of the decentralized, parent-choice nature of the program under NCLB, matching students with the most appropriate tutor is not always possible.

Time, effort, and public funding are wasted when students with special needs are not placed with providers or tutors who have the capacity to serve them. Conflicting day school and after-school instructional strategies can negatively impact the student's day school instruction and hinder the capabilities of the provider to meet the student's needs. In some cases, these issues lead to the student being transferred to another provider. However, it is often the case that parents are not aware of the differences in curriculum and instruction and the consequences (positive or negative) of those differences for their students' outcomes. Providers may max out their per-pupil allowance and hours to work with a student before they ever really understand the best strategies to help that student.

Overall, our quantitative and qualitative findings combined suggest that in many publicly funded, OST tutoring sessions, students are not getting enough hours of high-quality, differentiated instruction to produce significant gains in their learning. This is not a problem that will be resolved only by setting minimum hours standards for tutoring providers, given that invoiced hours do not equal quality instructional time.

CONCLUSION

Recent K–12 educational reform activity suggests that OST tutoring programs will persist as a mainstay intervention in federal, state, and district reform efforts. For

instance, the 21st Century Community Learning Centers (CCLCs) remain an important source of supplementary instruction for students in need, with federal appropriations of over $1 billion (as of 2011) for providing services to over 1.6 million students (After School Alliance, 2012). In addition, many districts with new freedom to design accountability programs are retaining tutoring as an important part of a systematic strategy to improve student outcomes. Tutoring also has potential to be a cornerstone in alternative models of schooling such as charter schools, where high-density tutoring has shown to generate significant gains in student achievement (Dobbie & Fryer, 2011).

This study has generated evidence that can be used by school districts as they pursue a broad spectrum of approaches to structuring (or redesigning) OST tutoring programs and to identifying new policy levers for implementation. Several of our study districts are already using this information to improve OST tutoring policy design and to develop new programs that are being launched following state waivers from NCLB. Even in the absence of a waiver, CPS instituted policies aimed at compelling providers to deliver more hours of tutoring via guidelines for using district space and the district provider's own rate setting (which has driven down market rates charged by other tutors). Following waiver approvals in Wisconsin and Minnesota, Milwaukee Public Schools and Minneapolis Public Schools now require tutoring providers to comply with maximum hourly rates and other requirements that will ensure students are offered a minimum of 40 hours of tutoring. Milwaukee has also taken actions to reduce provider direct costs of delivering tutoring (e.g., eliminating facility rental fees), and Minneapolis is establishing performance-based contracts with bonuses. And in response to the qualitative study findings that consistently showed discrepancies between providers' advertised length of tutoring sessions and actual instructional time, these school districts have developed new policies to tighten monitoring of programs and student attendance, including cross-checking student signatures on attendance records, assigning school-based coordinators responsibilities for supervision, and more regular, random monitoring of student participation in tutoring sessions by district staff.

School districts will also benefit from ongoing opportunities to describe and share strategies for addressing challenges with intraprovider variation in tutoring instruction quality and curriculum materials. For example, Minneapolis Public Schools is instituting more structure to ensure that OST tutoring providers will implement programming that provides Minnesota standard-based, focused, and developmental instruction in its new district tutoring program. Many of the 45 states and three territories that have adopted the Common Core State Standards (CCSS) are urging OST tutoring programs in their states to align their curricula to the CCSS. Creating and maintaining mechanisms for cross-district communications and sharing of effective policies, strategies, and practices has the potential to limit missteps or setbacks experienced with new policy development and to more rapidly improve services for students and their achievement outcomes.

For English language learners and students with disabilities, it is clear that immediate changes in policy and practice are needed. At a minimum, tutors delivering instruction to these student populations must have basic knowledge of how to effectively address students' unique needs. Under NCLB regulations, providers are allowed to hire tutors who lack the basic training and qualifications needed to serve students with special needs. NCLB fails to address alignment with other relevant federal policies such as IDEA or FERPA. Furthermore, confusion regarding responsibilities and lack of coordination around other laws that target these subgroups exacerbate the problems, such as precluding tutors from having necessary student educational information or delaying provision of tutoring services.

Although the generalizability of these study findings are enhanced by the cross-district research design, we still have a relatively small sample of the many districts

in the United States where OST tutoring interventions are being implemented, and the considerable variation in tutoring providers and contexts may limit their applicability. Our empirical methods are also limited by the assumptions we are required to make in the absence of random assignment of students to what is a voluntary educational intervention. While we believe that the complementary findings from our qualitative and quantitative investigations and the triangulation of a number of qualitative and quantitative methods in the data analysis and interpretation of our results strengthens their credibility, we suggest that our study findings be applied with care and attention to state, district, provider, and student contexts, and to new developments on the OST tutoring program and research frontiers.

CAROLYN J. HEINRICH is a Professor of Public Affairs and affiliated Professor of Economics, University of Texas at Austin, 2315 Red River Street, P.O. Box Y, Austin, TX 78713. (E-mail: cheinrich@austin.utexas.edu.)

PATRICIA BURCH is an Associate Professor, Rossier School of Education at University of Southern California, 3470 Trousdale Parkway, Los Angeles, CA 90089-4038. (E-mail: pburch@usc.edu.)

ANNALEE GOOD is an Associate Researcher, University of Wisconsin-Madison, 1025 W. Johnson Street, Madison, WI 53706. (E-mail: aggood@wisc.edu.)

RUDY ACOSTA is a Research Assistant, PhD Candidate, Rossier School of Education at University of Southern California, 3470 Trousdale Parkway, Los Angeles, CA 90089-4038. (E-mail: acostar@usc.edu.)

HUIPING CHENG is an Associate Researcher, University of Wisconsin-Madison, 1025 W. Johnson Street, Madison, WI 53706. (E-mail: hcheng6@wisc.edu.)

MARCUS DILLENDER is an Economist, W.E. Upjohn Institute for Employment Research, 300 S. Westnedge Avenue, Kalamazoo, MI 49007-4686. (E-mail: dillender@upjohn.org.)

CHRISTI KIRSHBAUM is a Research Associate, University of Texas at Austin, 114 Shady Trails Pass, Cedar Park, TX 78613. (E-mail: Christi.kirshbaum@utexas.edu.)

HIREN NISAR is a Senior Analyst, PhD Economist, Abt Associates Inc., 6005 4th Street, NW, Washington, DC 20011. (E-mail: Hiren_Nisar@abtassoc.com/hirenius@gmail.com.)

MARY STEWART is an Associate Researcher, University of Wisconsin-Madison, 310 W. Kenwood Drive, Bloomington IN 47404. (E-mail: Mollymarys@gmail.com.)

REFERENCES

After School Alliance. (2012). 21st Century Community Learning Centers Federal Afterschool Initiative. Retrieved July 25, 2012, from http://www.afterschoolalliance.org.

Barnhart, M. (2011). The impact of participation in supplemental education services on student achievement: 2009–10. Los Angeles Unified School District Research Unit No. 379. Los Angeles, CA: Los Angeles Unified School District.

Beckett, M., Borman, G., Capizzano, J., Parsley, D., Ross, S., Schirm, A., & Taylor, J. (2009). Structuring out-of-school time to improve academic achievement: A practice guide NCEE #2009–012. Washington, DC: National Center for Education Evaluation and Regional

Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved August 5, 2012, from http://ies.ed.gov/ncee/wwc/publications/practiceguides.

Bifulco, R. (2012). Can nonexperimental estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. Journal of Policy Analysis and Management, 31, 729–751.

Black, A. R., Doolittle, F., Zhu, P., Unterman, R., & Grossman, J. B. (2008). The evaluation of enhanced academic instruction in after-school programs: Findings after the first year of implementation. NCEE 2008–4021. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Burch, P. (2009). Hidden markets: New education privatization. New York, NY: Routledge.

Burch, P., & Good, G. A. (in press). Equal Scrutiny: Privatization and Accountability in Digital Education. Cambridge, MA: Harvard Education Press.

Burch, P., Heinrich, C. J., Good, A., & Stewart, M. (2011, February). Equal access to quality in federally mandated tutoring: Preliminary findings of a multisite study of supplemental educational services. Working paper presented at the 2011 Sociology of Education Conference, Monterey, CA.

Cook, T., Shadish, W., & Wong, V. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. Journal of Policy Analysis and Management, 27, 724–750.

Deke, J., Dragoset, L., Bogen, K., & Gill, B. (2012). Impacts of Title I supplemental educational services on student achievement. NCEE 2012-4053. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Dobbie, W., & Fryer, R. (2011). Getting beneath the veil of effective schools: Evidence from New York City. Cambridge, MA: Education Innovation library.

Durlak, R., & Weissberg, R. (2007). The impact of after-school programs that promote personal and social skills. Chicago, IL: CASEL.

Dynarski, M., James-Burdumy, S., Moore, M., Rosenberg, L., Deke, J., & Mansfield, W. (2004). When schools stay open late: The national evaluation of the 21st Century Community Learning Centers Program: New findings. Washington, DC: U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, U.S. Government Printing Office.

Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. Journal of Educational Psychology, 92, 605–619.

Farkas, G. (2000). Tutoring for low-income children via vouchers to their parents. Journal of Policy Analysis and Management, 19, 143–145.

Farkas, G., & Durham, R. (2007). The role of tutoring in standards-based reform. In A. Gamoran (Ed.), Standards-based reform and the poverty gap (pp. 201–228). Washington, DC: Brookings Institution.

Fryer, R. G. (2012). Injecting successful charter school strategies into traditional public schools: Early results from an experiment in Houston. Working paper. Cambridge, MA: Harvard University.

Heinrich, C. J., Meyer, R. H., & Whitten, G. (2010). Supplemental education services under No Child Left Behind: Who signs up, and what do they gain? Educational Evaluation and Policy Analysis, 32, 273–298.

Heinrich, C. J., & Nisar, H. (2013). The efficacy of private sector providers in improving public educational outcomes. American Educational Research Journal, 50, 856–894.

Heistad, D. (2007). Evaluation of supplemental education services in Minneapolis Public Schools: An application of matched sample statistical design. Minneapolis, MN: Office of Research, Evaluation and Assessment, Minneapolis Public Schools.

Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. In A. Gelman & X. L. Meng (Eds.), Applied Bayesian modeling and causal inference from incomplete-data perspectives (pp. 73–84). West Sussex, England: Wiley InterScience.

Jones, C. (2009). The 2009 Supplemental Educational Services program: Year 4 summative evaluation. Chicago, IL: Chicago Public Schools Office of Extended Learning Opportunities and Office Research, Evaluation, and Accountability.

Lauer, P., Akiba, M., Wilkerson, S., Apthorp, H., Snow, D., & Martin-Glenn, M. (2006). Out-of-school-time programs: A meta-analysis of effects for at-risk students. Review of Educational Research, 76, 275–313.

Little, P., Wimer, C., & Weiss, H. (2008). Programs in the 21st century: Their potential and what it takes to achieve it. Cambridge, MA: Harvard Family Research Project.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping. A meta-analysis. Review of Educational Research, 66, 423–458.

Pianta, R., & Hamre, B. (2009). Conceptualization, measurement and improvement of classroom processes: Standardized observation can leverage capacity. Educational Researcher, 38, 109–119.

Rickles, J. H., & Barnhart, M. K. (2007). The impact of supplemental educational services participation on student achievement: 200506. Report, Pub. No. 352. Los Angeles, CA: Los Angeles Unified School District Program Evaluation and Research Branch, Planning, Assessment and Research Division.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41–55.

Springer, M. G., Pepper, M. J., & Ghosh-Dastidar, B. (2009). Supplemental educational services and student test score gains: Evidence from a large, urban school district. Working paper. Nashville, TN: Vanderbilt University.

U.S. Department of Education. (2005). No Child Left Behind: Supplemental educational services non-regulatory guidance (Final Guidance). Washington, DC: Author.

U.S. Department of Education. (2009). Supplemental educational services: Non-regulatory guidance. Retrieved November 17, 2011, from http://www2.ed.gov/policy/elsec/guid/suppsvcsguid.doc.

Vandell, D. L., Reisner, E. R., Brown, B. B., Dadisman, K., Pierce, K. M., Lee, D., & Pechman, E. M. (2005). The study of promising after-school programs: Examination of intermediate outcomes in year 2. Madison, WI: Wisconsin Center for Education Research.

Vandell, D., Reisner, E., & Pierce, K. (2007). Outcomes linked to high-quality afterschool programs: Longitudinal findings from the study of promising practices. Irvine, CA: University of California and Washington, DC: Policy Studies Associates.

Zimmer, R., Gill, B., Razquin, P., Booker, K., Lockwood, J. R., Vernez, G., Birman, B. F., Garet, M. S., & O'Day, J. (2007). State and local implementation of the No Child Left Behind Act: Volume I—Title I school choice, supplemental educational services, and student achievement. Washington, DC: U. S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

Zimmer, R., Hamilton, L., & Christina, R. (2010). After-school tutoring in the context of No Child Left Behind: Effectiveness of two programs in the Pittsburgh Public Schools (PPS). Economics of Education Review, 29, 18–28.

APPENDIX A: QUALITATIVE METHODS DETAILS

As described in the main text, the qualitative analysis in this paper draws on qualitative data collected over the 2009–2010 to 2011–2012 school years. Our sample of 23 SES providers across four districts includes 12 providers described as having school-based services, five as community-based, and 11 as home-based. A number of providers described themselves as providing tutoring in more than one location. Four of the providers in the qualitative sample rely solely on digital tutoring platforms. The qualitative dataset used in this paper includes observations of full tutoring sessions ($n = 123$), interviews with provider administrators ($n = 55$), interviews with tutoring staff ($n = 69$), interviews with district and state administrators ($n = 31$), parent focus groups ($n = 155$), and archival document analysis.

Table A1 presents a breakdown of the number of OST tutoring session observations, as well as the tutoring formats and locations observed. In some cases, district policy affected the types of tutoring we could observe. For example, in-home tutoring was not permitted by CPS; therefore, we do not have any observations of home-based tutoring sessions in Chicago. The purpose of these observations was to gain insight on variability within or trends across OST tutoring context as a whole, and tutoring formats in particular. We do not make assertions about specific providers based solely on observation data. We made every effort to observe tutoring sessions with the tutors that we had also interviewed in order to provide multiple data points on the same instructional experience.

We conducted regular reliability trainings with the qualitative research team throughout the project to ensure consistency in observation ratings. In each session the research team rated the same video segment of an instructional session and went through each indicator to compare ratings. Validity of the instrument was ensured by the development process, whereas its structure and content was based on well-tested, existing observation instruments for OST tutoring, existing literature on the best practices for OST tutoring, and the theory of action in the supplemental educational services policy. We continued to test and refine the data collection process as the study progressed. A copy of the observation instrument is available at http://www.sesiq2.wceruw.org.

Table A2 shows the specific numbers of interviews conducted for this project. Our qualitative work includes annual interviews with program directors of each tutoring provider. The purpose of these provider director interviews was to identify important program attributes that may not otherwise be observed in quantitative data and to develop measures for exploring their associations with program effects. Director interviews focused on recruitment and retention strategies; curriculum alignment efforts; staff training; communication strategies with the district, school, and families; and the guiding principles of their program and diagnostic strategies. We asked directors to identify five to 10 tutors working in the schools in which they have students enrolled. We then selected tutors to interview from within this list to

Table A1. Observations of full tutoring sessions in 2009 to 2010, 2010 to 2011, and 2011 to 2012 school years.

	Home	School	Community	Digital
Chicago	0	22	1	1
Dallas	0	15	3	8
Milwaukee	10	22	0	6
Minneapolis	10	2	11	12
Total (123)	20	61	15	27

Table A2. Interviews/focus groups in 2009 to 2010, 2010 to 2011, and 2011 to 2012 school years.

	Provider and admin	Provider and tutor	District admin	State admin	Parent focus group
Chicago	12	16	6	2	16
Dallas	10	11	7	1	45
Milwaukee	15	18	5	1	33
Minneapolis	18	24	7	2	61
Total	55	69	25	6	155

Table A3. Documents collected in 2009 to 2012 field work.

Policy documents	Curriculum materials/assessments
• State policies regarding incentives • Legal complaints • Internal/external evaluations • Provider contact lists • Sample invoices • Individualized learning plans (template and examples stripped of student identifying information) required of providers by some districts • State explanation of monitoring process • Sample contracts • Completed and evaluated applications to the state, including state rubric	• Formal curriculum • Copy of lessons plans • Teacher guide • Sample worksheets • PowerPoint presentations that lay out structure of online curriculum • Home grown and commercially prepared assessments used by providers pre/post-intervention • Software curriculum used in nonlive tutoring program
Other provider materials	**Communication**
• Instructor attendance logs • Marketing materials • Student attendance records/log forms • Research base for curriculum and instruction • Ongoing progress monitoring results for participating students (anonymous)	• District and school • District and provider correspondence • Staff evaluation forms • Provider and parent communications • Demo training CD • Tutor and parent e-mails and letters • Tutor and school teacher e-mails, letters, and progress reports. • School coordinators and parents

minimize bias. Tutor interviews focused on areas identified in district interviews as well as particular adaptations that the tutor uses in practice.

We also conducted structured interviews with district and state administrators focused on issues such as interorganizational coordination, organizational capacity, interaction with policy requirements, and with other districts in support of the use of evidence in policy and program decisions and program improvement.

One criterion for focus groups with parents and guardians was that the student (of the invited parent) would have been offered the opportunity to participate in OST tutoring. Parent focus groups examined school-, district-, and provider-level attributes that might be associated with program effects on student achievement.

Lastly, see Table A3 for a detailed list of the types of archival documents collected to further triangulate the qualitative data on instructional formats, as well as program structure and communication strategies.

APPENDIX B: QUANTITATIVE METHODS AND ANALYSIS DETAILS

ALTERNATIVE MODEL SPECIFICATIONS TESTED

In estimating average effects of OST tutoring, we adjusted for student selection into tutoring using three alternative strategies.

Value-Added Model

The value-added strategy specified in equation (B.1), our main specification, allows us to control for other classroom and school interventions that are fixed over time, while identifying provider characteristics. We estimate the following equation:

$$A_{jst} - A_{jst-1} = \alpha \mathrm{OST}_{jt} + \beta X_{jt-1} + \pi_s + \mu_{gt} + E_{jst} \tag{B.1}$$

where A_{jst} is the achievement of student j attending school s in year t; OST_{jt} is an indicator function if student j attended tutoring in year t; X_{jt-1} are student characteristics that include student demographics, percent absent in prior year, retained in prior year, and attended tutoring in the prior year; π_s is school fixed effect; μ_{gt} are grade by year fixed effects; and E_{jst} is the random error term. Identification in this specification comes from the average gain in student achievement after controlling for student characteristics and school and grade year effects.

Student Fixed Effects Model

The value-added model assumes that selection depends on observed student characteristics. Hence, controlling for them allows us to deal with self-selection. However, if selection is on some unobserved or unmeasured characteristics of the students, then a value-added strategy could still lead to biased results. The student fixed effects model controls for all time-invariant characteristics of a student, including those that are not observed or measured. The following model of an educational production differs from equation (B.1) in that it includes student fixed effects (δ_j) instead of school fixed effects,

$$A_{jst} = \alpha \mathrm{OST}_{jt} + \beta X_{jt-1} + \delta_j + \mu_{gt} + E_{jst}. \tag{B.2}$$

When we take the first difference of equation (B.2), we eliminate the student fixed effect (δ_j), and the model estimates the average difference between the gains made by students attending OST tutoring with the gains made by similar students who were likewise eligible for services. This formulation imposes some restrictions (or assumptions) that are important to note. First, the impact of students' prior experience does not deteriorate over time. This implies, for example, that the effect of the quality of kindergarten has the same impact on student achievement no matter the grade. The second assumption is that the unobserved effect of attending tutoring only affects the level, but not the rate of growth in student achievement. A concern with this restriction is that if students with lower growth are more likely to choose to attend OST tutoring, then this type of selection may bias the estimates obtained from a gains model.

In order to relax this restriction, the following equation is estimated,

$$A_{jst} - A_{jst-1} = \alpha \text{OST}_{jt} + \beta X_{jt-1} + \delta_j + \mu_{gt} + E_{jst}. \quad \text{(B.3)}$$

This approach to estimating the fixed effects model controls for any unobserved differences between students that are constant across time. The estimation of this model requires a first difference of equation (B.3) and therefore needs three or more observations for each student. As students self-select into the tutoring program, we deal with this by using the gain scores made by the same student in the prior year. Identification of the average impact of tutoring in this model comes from students who participate in one or more, but not all years. If these students differ in systematic ways from all students who receive tutoring, then the estimator gives a *local* effect (specific to students with these characteristics) instead of an average effect.

School and Student Fixed Effects Model

The base model for this estimation strategy is the combination of the two above methods. A school fixed effect (π_s) is added to equation (B.3), which gives

$$A_{jst} - A_{jst-1} = \alpha \text{OST}_{jt} + \beta X_{jt-1} + \pi_s + \delta_j + \mu_{gt} + E_{jst}. \quad \text{(B.4)}$$

The inclusion of school fixed effects facilitates controlling for time-invariant school characteristics such as average school test scores, neighborhood attributes, parental involvement in the school, and peer composition, to the extent these are unchanging over time. The inclusion of student fixed effects effectively controls for student ability and other time-invariant student characteristics.

For brevity, we present the estimates of average impacts of OST tutoring on student math and reading achievement from these alternative model specifications for one school district only (CPS). As shown below in Table B1, the results are robust across all specifications. As a result, we only include the school value added results in the main text of the paper. Results for other sites and specifications are available from the authors upon request.

ESTIMATES WITH ALTERNATIVE METHODS FOR CONTROLLING FOR STUDENTS' PRIOR ACHIEVEMENT (TEST SCORES)

As discussed above and in the main text, the estimates are the α coefficients from estimating the value-added model with school fixed effects (equation B.1 above): $A_{jst} - A_{jst-1} = \alpha \text{OST}_{jt} + \beta X_{jt-1} + \pi_s + \mu_{gt} + E_{jst.}$ The purpose of taking into account A_{jst-1} in the main estimating equation is that students with different abilities may be more or less likely to sign up for OST tutoring, and accounting for pretest measures has been shown to be a reliable way to deal with this self-selection into treatment (Bifulco, 2012; Cook, Shadish, & Wong, 2008). In this section we consider alternative approaches to controlling for prior test performance, focusing on the 2011 to 2012 estimates of OST tutoring impacts, which allows us to take advantage of having multiple (prior year) test scores. For ease of comparing the estimates from alternative specifications, the various estimates and their 95 percent confidence intervals are presented graphically for each district in Figures B1 through B8.

As a baseline, we consider models that do not use previous test scores to account for selection into OST. To do so, we estimate models of the following form:

$$A_{jst} = \alpha \text{OST}_{jt} + \beta X_{jt-1} + \pi_s + \mu_{gt} + E_{jst}. \quad \text{(B.5)}$$

Table B1. Average impacts of OST tutoring on reading and math achievement (gains) via alternative estimation strategies in Chicago Public Schools.

	Reading									
	School value-added model						Student fixed effects model		School and student fixed effects model	
	Year 2008 to 2009		Year 2009 to 2010		Year 2010 to 2011					
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
Attended OST	0.043	0.006	0.094	0.009	0.075	0.009	0.085	0.024	0.087	0.024
Number of observations	61,171		63,506		80,510		124,677		124,677	
Number of schools	227		454		302		458		458	
Number of students	61,171		63,506		80,510		83,945		83,945	

	Math									
	School value-added model						Student fixed effects model		School and student fixed effects model	
	Year 2008 to 2009		Year 2008 to 2009 Year 2009 to 2010		Year 2010 to 2011					
	Coef	SE			Coef.	SE	Coef.	SE	Coef.	SE
Attended OST	0.046	0.005	0.053	0.008	0.064	0.009	0.054	0.021	0.055	0.021
Number of observations	61,464		63,773		80,614		124,059		124,059	
Number of schools	227		455		302		458		458	
Number of students	61,464		63,773		80,614		83,579		83,579	

The α coefficients from equation B.5 are the first coefficients shown in Figures B1 through B8 and are labeled *No Pretest*. For comparison, we also show the original estimates from the paper in Figures B1 through B8. They are labeled as the *Change* estimates and are the α coefficients obtained from estimating our main model (B1). For Chicago, Dallas, and Milwaukee, the No Pretest coefficients from equation B.5 appear to be smaller than the main specification coefficients, implying that there is negative selection into OST based on ability. For Minneapolis Public Schools, the No Pretest estimates appear to be larger than the other estimates, which suggest that there may be positive selection into OST based on ability on this district.

Our main estimates use the student's previous year's test score to create a change measure, which we then use as the dependent variable to estimate the effect of OST. Another specification that is also common in the literature is to instead use the posttest standardized score as the dependent variable and to control for the student's previous year test score on the right-hand side. The third set of estimates shown in Figures B1 through B8 are the α coefficients from estimating models of this form:

$$A_{jst} = A_{jst-1} + \alpha \text{OST}_{jt} + \beta X_{jt-1} + \pi_s + \mu_{gt} + E_{jst}. \tag{B.6}$$

In almost all cases, these *Level on Previous Level* estimates fall within the 95 percent confidence interval of the original estimates. While the signs on these estimates differ from the signs of the Change estimates when estimating the effect of OST tutoring on math achievement in Milwaukee Public Schools, these estimated effects

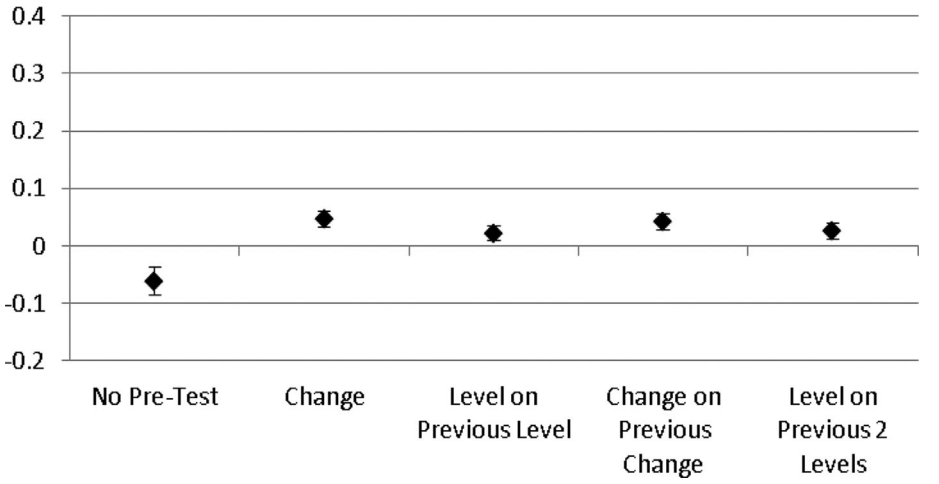


Figure B1. Chicago Math Estimates.

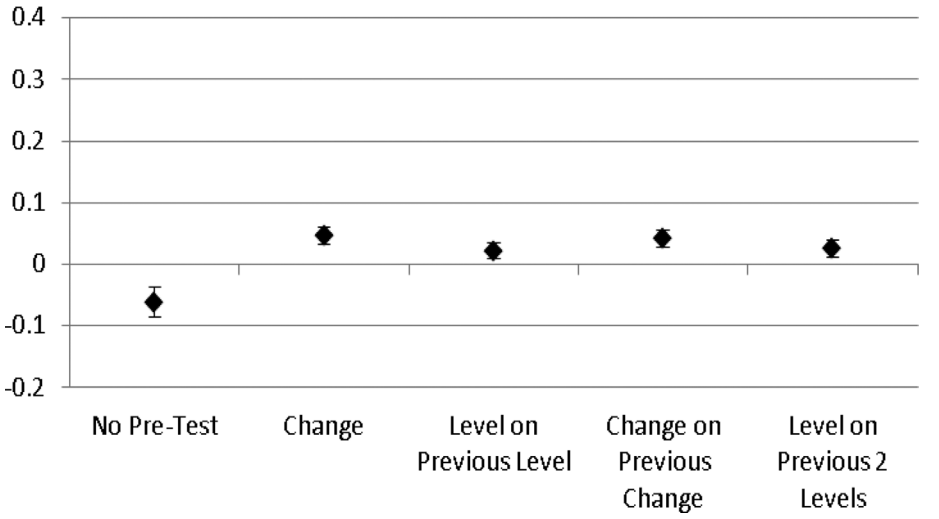


Figure B2. Chicago Reading Estimates.

of OST on math achievement in Milwaukee are statistically indistinguishable from zero and from each other.

These results suggest that accounting for selection into OST tutoring is important. A potential concern is that using one year's prior score may be too noisy to fully account for selection. The fourth set of estimates we show in Figures B1 through B8 controls for the student's previous year's score change, which is equivalent to estimating models of the following form:

$$A_{jst} - A_{jst-1} = \Delta A_{jst-1} + \alpha \text{OST}_{jt} + \beta X_{jt-1} + \pi_s + \mu_{gt} + E_{jst} \quad \text{(B.7)}$$

where $\Delta A_{jst-1} = A_{jst-1} - A_{jst-2}$. In all cases, these *Change on Previous Change* estimates are almost identical to the estimates from equation (B.1), which suggests that using multiple score measures does not do a better job of controlling for selection than using only one previous score.

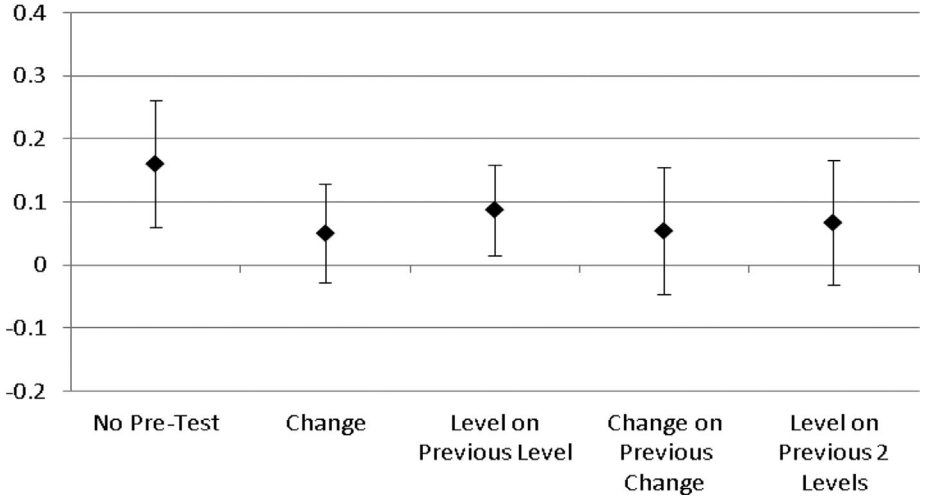


Figure B3. Minneapolis Math Estimates.

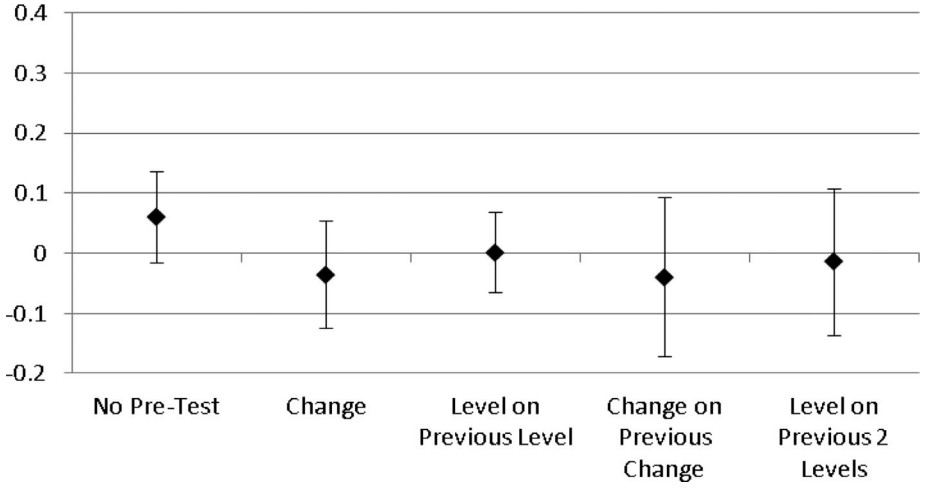


Figure B4. Minneapolis Reading Estimates.

Finally, we also test a specification with an additional test score measure on the right-hand side of the equation and report these estimates in Figures B1 through B8:

$$A_{jst} = A_{jst-1} + A_{jst-2} + \alpha \mathrm{OST}_{jt} + \beta X_{jt-1} + \pi_s + \mu_{gt} + E_{jst}. \tag{B.8}$$

For all districts, the coefficients from this specification appear to be very similar to those of model B6 that controls for just one prior year test score.

The results presented here thus imply that accounting for selection into OST tutoring is important. Whether we account for selection by creating a change measure or by controlling for the lagged test score, we find similar results. Similarly, using multiple prior test scores to account for selection changes the results very little, suggesting using one pretest measure is likely sufficient.

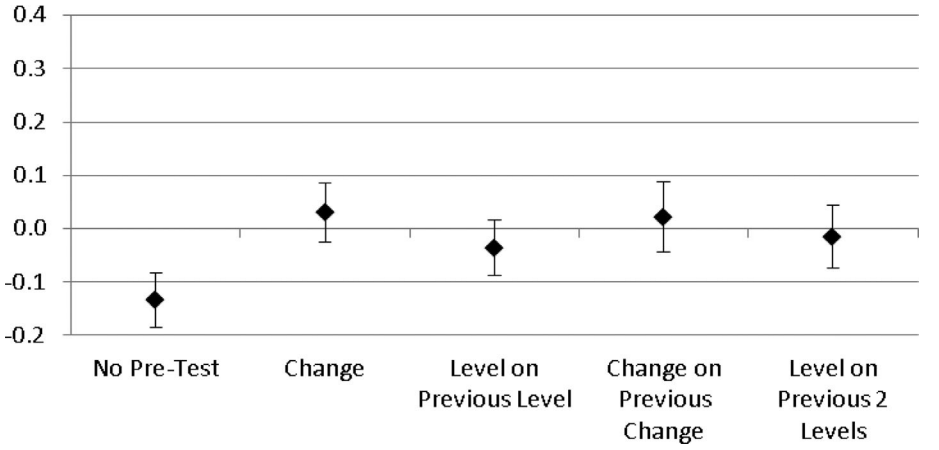


Figure B5. Milwaukee Math Estimates.

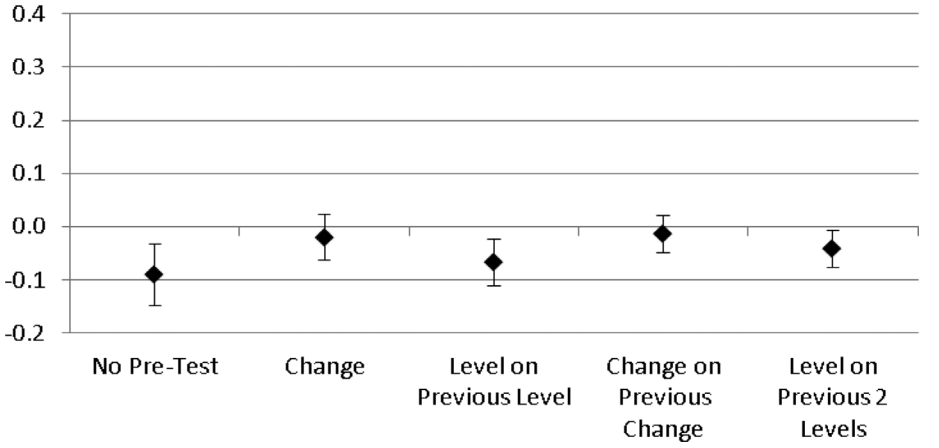


Figure B6. Milwaukee Reading Estimates.

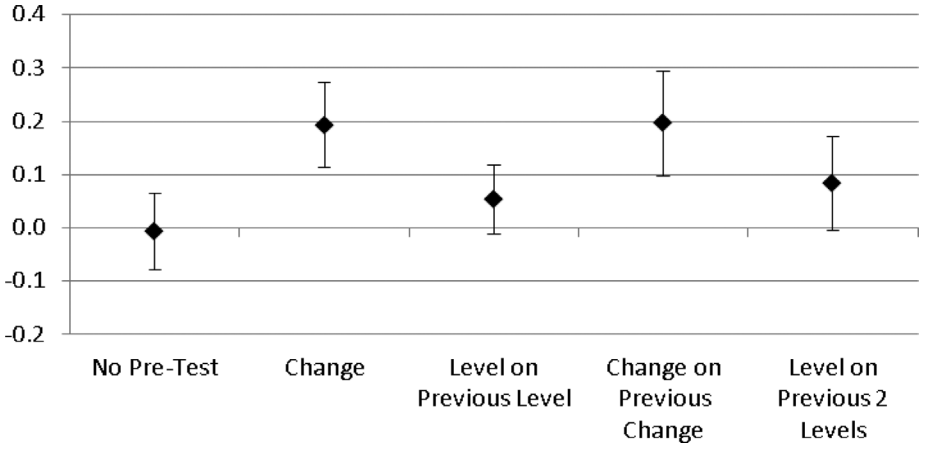


Figure B7. Dallas Math Estimates.

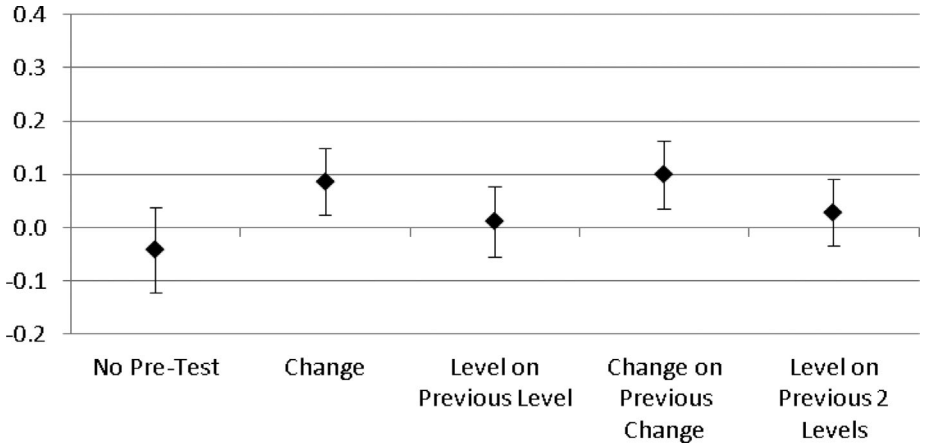


Figure B8. Dallas Reading Estimates.

GPS MATCHING DETAILS

As indicated in the main text, we follow the notation for GPS matching formalized by Hirano and Imbens (2004). They define $Y_i(t)$ as the set of potential outcomes of treatment $t \in \Im$, (i.e., t is an element in the set, $\Im$), where $\Im$ may be an interval of a continuous treatment. For each student i, we observe a vector of covariates, X_i (that predict take-up of the treatment); the level of the treatment, T_i, that the student, i, actually receives, and the potential outcome associated with the level of treatment received, $Y_i = Y_i(T_i)$.

The (weak) unconfoundedness assumption states that conditional on observed covariates, the level of treatment received (T_i) is independent of the potential outcome, $Y_i\ (t) \perp T_i \mid X_i$ for all $t \in \Im$. In other words we assume there is no systematic selection into levels of treatment based on unobservable characteristics.

The conditional density, that is, conditional on pretreatment covariates, of the treatment is $r(t, x) = f_{T\mid X}\ (t \mid x)$, and the GPS is therefore defined as the conditional density of receiving a particular level of the treatment, $t = T$: $R = r(T, X)$.

Similar to the binary treatment (PSM) case, the GPS balances the covariates *within strata* defined by values of the GPS, so that the probability that $t = T$ does not depend on the value of X (and assignment to treatment *levels* is unconfounded). As Hirano and Imbens (2004) show, this allows the estimation of the average dose–response function, $\mu(t) = E[Y_i\ (t)]$, using the GPS to remove selection bias.

After estimating the GPS, the next step is to estimate the conditional expectation of the outcome (Y_i) as a function of the treatment level, T, and the GPS, R: $\beta(t, r) = E[Y|T = t, R = r]$. The regression function, $\beta(t,r)$, represents the average potential outcome for the strata defined by $r(T,X) = r$, but it does not facilitate causal comparisons across different levels of treatment. That is, one cannot directly compare outcome values for different treatment levels to obtain the causal difference in the outcome of receiving one treatment level versus another.

A second step is required to estimate the dose–response function at each particular level of the treatment. This is implemented by averaging the conditional means, $\beta(t,r)$, over the distribution of the GPS, $r(t,X)$, that is, for each level of the treatment:

$$\mu(t) = E[\beta(t, r(t, X))],$$

where $\mu(t)$ corresponds to the value of the dose–response function for treatment value t, and when compared to another treatment level, does have a causal interpretation.

As in propensity score matching, it is also important to assess the balance of the covariates following GPS estimation. We follow the approach of Agüero et al. (2007), who defined three different treatment terciles of the treatment variable and tested whether the mean value of the covariates were the same for the observations in the different treatment terciles. They then investigated whether the covariates were better balanced after conditioning on the estimated GPS. We apply four cut points to the treatment variable, which generates five treatment intervals, and we correspondingly check for balance of covariates within them.

We implement GPS with the Stata *gpscore* and *doseresponse* commands (that need to be installed in Stata). Students who did not receive any OST tutoring are excluded from estimation; this procedure is for estimating the effects of treatment intensity (dose–response) and not for comparing those with and without treatment. Several tests are built into the code shown below, including a test for normality of the disturbances and a test that the conditional mean of the pretreatment variables given the GPS is not different between units that belong to a particular treatment interval and units that belong to all other treatment intervals (i.e., the balancing test). Cut points for the treatment variable were defined by quartiles of the treatment distribution, as well as the levels of treatment for which the dose–response function estimates the average potential outcome (10 to 80 hours). Bootstrapped standard errors are also requested. Different transformations for the treatment variable can also be specified; we specify a linear model.

Stata Code

```
matrix define tp = (10\15\20\25\30\35\40\45\50\55\60\65\70\75\80)
qui generate cut4 = 32 if tseshrs_attended< = 32
qui replace cut4 = 42 if tseshrs_attended>32 & tseshrs_attended< = 42
qui replace cut4 = 54 if tseshrs_attended>42 & tseshrs_attended< = 54
qui replace cut4 = 55 if tseshrs_attended>54
doseresponse female ell frl sped white asian hispanic otherrace
attendedses retained percentabsent_ly dum_grade2-dum_grade6,
outcome(mathgain) t(tseshrs_attended) gpscore(gscore)
predict(SES_fitted) sigma(g_std) cutpoints(cut4) index(p50)
nq_gps(5) dose_response(dose_response) reg_type_t(linear)
reg_type_gps(linear) bootstrap(yes) boot_reps(50) tpoints(tp)
delta(0) analysis(yes) interaction(1) detail
```

REFERENCES

Agüero, J., Carter, M., & Woolard, I. (2007, September). The impact of unconditional cash transfers on nutrition: The South African Child Support Grant. International Poverty Centre Working Paper No. 39. Brasilia, Brazil: International Poverty Centre.

Bifulco, R. (2012). Can nonexperimental estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. Journal of Policy Analysis and Management, 31, 729–751.

Cook, T., Shadish, W., & Wong, V. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. Journal of Policy Analysis and Management, 27, 724–750.

Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. In A. Gelman & X. L. Meng (Eds.), Applied Bayesian modeling and causal inference from incomplete-data perspectives (pp. 73–84). West Sussex, England: Wiley InterScience.