May 2006

Education Policy Analysis Archives 14/12

Arizona State University

University of South Florida

EDUCATION POLICY ANALYSIS ARCHIVES

A peer-reviewed scholarly journal
Editor: Sherman Dorn
College of Education
University of South Florida

Volume 14 Number 12 May 1, 2006 ISSN 1068–2341

Gathering Evidence on an After-School Supplemental Instruction Program: Design Challenges and Early Findings in Light of NCLB[1]

Madhabi Chatterji
Teachers College, Columbia University

Young Ae Kwon
Kwon Learning Center

Clarice Sng
Teachers College, Columbia University

Citation: Chatterji, M., Kwon, Y.A., & Sng, C. (2006). Gathering evidence on an after-school supplemental instruction program: Design challenges and early findings in light of NCLB. *Education Policy Analysis Archives, 14*(12). Retrieved [date] from http://epaa.asu.edu/epaa/v14n12/.

Abstract
The No Child Left Behind (NCLB) Act of 2001 requires that public schools adopt research-supported programs and practices, with a strong recommendation for randomized controlled trials (RCTs) as the "gold standard" for scientific rigor in empirical research. Within that policy framework, this paper compares the relative utility of federally-recommended RCT versus the demonstrated *extended term mixed-method* (ETMM) designs as options for monitoring effects of novel

[1] The empirical study embedded in this paper was conducted at the request of the supplemental instruction program provider. The first author thanks the program providers and school leaders for their support and facilitation during the conduct of the study. An earlier version of this paper was presented at the Annual Meeting of the American Educational Research Association held at San Diego, CA on April 13, 2004. Names are not released to honor client confidentiality.

programs in real-time field settings. Guided by the program's theory of action, a year-long, two-phase study was conducted to monitor the context, processes and early outcomes of an after-school supplemental program in a New York elementary school. In both phases, the design combined a matched-groups, quasi-experiment with qualitative classroom observations and descriptive surveys. Early findings showed some positive, albeit "gross" program effects. Although findings are tentative, the ETMM approach enhanced interpretations by shedding light on relevant environmental variables, causes for program instabilities and sample attrition, and factors affecting treatment fidelity and scaling-up of the program beyond the pilot year.
Keywords: research evidence; supplemental instructional programs; rigorous evaluation methods.

The No Child Left Behind (NCLB) Act of 2001 requires that public institutions adopt research-supported programs, practices and policies, with a strong recommendation for the use of randomized controlled trials (RCT) as the "gold standard" for attaining scientific rigor in empirical research efforts (U.S. Department of Education, 2003). Within that policy framework, this paper compares the relative utility of federally-recommended RCT versus the demonstrated *extended term mixed-method* (ETMM) design (Chatterji, 2005) as options for monitoring effects of novel programs in real-time field settings. To demonstrate the merits and demerits of the alternate ETMM approach, this article details the design concepts and empirical procedures employed to monitor early processes and effects of a supplemental program in reading and mathematics, as implemented in one elementary school in New York City. Design challenges that were faced along the way and modifications made to the original design are discussed against the body of information that was obtained on conclusion of the two-phase, mixed-method investigation. To allow for a comparative appraisal of the utility of the demonstrated ETMM approach against RCT by readers, a federally-funded national evaluation of another supplemental instruction program, the 21st Century Community Learning Centers (21st CCLC) (Mathematica Policy Research, Inc. & Decision Information Resources, Inc., 2003) is used as a benchmark for discussion.

Theoretical Framework

As a preamble to a detailed presentation of the specific ETMM application with the New York City supplemental program evaluation, the article begins with a discussion of three topics. First, it examines the NCLB context for schools to adopt programs and practices supported by research evidence, the meaning of "scientific rigor" as given in federal documents, and the difficulties in implementing sound experiments in field settings. Second, it describes a concurrent federal recommendation that emerged under the auspices of the NCLB, namely, that failing schools should utilize supplemental instruction services and extended day schooling to improve student achievement outcomes. Lastly, it describes the 21st CCLC study, which was supported by federal funds and where evaluation researchers attempted to implement RCTs on a national scale.

Federal Mandate for "Scientific Rigor" and Difficulties in Mounting Rigorous Experiments

Soon after the passage of NCLB in 2001, the Coalition for Evidence-based Policy under the DOE's Institute of Education Sciences (IES) released formal guidelines on identifying and implementing evidence-based practices in K-12 systems. Calling on educational practitioners to comply with the NCLB mandate for using "scientifically-based research" to guide their decisions about programs and interventions to implement (U.S. Department of Education, 2003), the document identified randomized controlled trials (RCTs) as the "gold standard" for obtaining strong and rigorous evidence on the effects of field-based programs and interventions. RCTs were defined as empirical studies that measure comparative effects of an intervention by randomly assigning individuals to the new program and to a control condition.

Several providers, independent researchers and research agencies have since made valiant attempts to respond to the federal requirement for executing randomized experiments on educational and other programs in public institutions. However, barriers in field settings have been numerous.

Due to organizational, political, and day-to-day operational complexities in schools and districts, true experiments are difficult to mount—whether in the case of supplemental or mainstream school innovations (see Cook, 2002, for a list of barriers). Quasi-experimental, time-series, and regression discontinuity designs have been suggested as alternatives for making generalized causal inferences on educational programs (Shadish, Cook & Campbell, 2002). Some quasi-experimental designs have limited applicability to particular classes of problems (for example, regression discontinuity approaches are best applied when differential placement of subjects is a part of the treatment program design). All experimental designs, however, tend to emphasize *outcomes*. Further, they assume that "treatments" can easily be standardized in and across field sites, and that effects can be fairly measured and compared once "treatment fidelity" is obtained and inter-pupil differences equalized in treatment and control groups, holding all else constant in the environment as long as the experiment continues.

In actuality, it is not easy to gather definitive empirical evidence of treatment fidelity in typical school settings, because educational treatments are not singular, narrowly-scripted entities. Even when gathered, qualitative differences in day-to-day operational definitions of a program make it difficult to draw conclusive causal inferences between a program and measured outcomes, particularly when a program is new. Further, while effective random assignment of subjects (the *sine qua non* of the "true" experiment) may statistically equalize pre-existing differences in pupils, the procedure cannot erase interfering effects of potential contextual contaminants. Multiple and often dissimilar initiatives are commonly in operation in open, complex, hierarchical systems that schools represent, all often targeting the same outcomes in the same groups of children. Control conditions often overlap and are not markedly dissimilar in operation from the treatments in early implementation phases.

In cases where similar groups of pupils can be assigned to treatment and control conditions and the treatment delivered in a stable manner, two added sets of factors *must* be taken into consideration when designing school-based studies on supplemental or mainstream services. The first deals with the *time* needed for the critical, operational components of a program to settle down and for the program to take shape at a given site. The second deals with *environmental dynamics* during the course of a study that may alter the operational definitions of treatment, control, and other confounding conditions in complex organizations. Because they are added instructional opportunities appended to an array of regular-day initiatives, the design challenges are particularly acute when studying effects of supplemental instruction programs on student achievement levels.

Chatterji (2004, 2005) thus recently asserted that comparative experiments by themselves are inadequate designs for studying school-based initiatives and proposed broader ETMM designs as an alternative. ETMM designs complement experimental designs with other methods, and use a phased approach in executing the research in order to better study environmental, treatment and control variables *in situ*, while allowing the program to take hold.

Historically, methodological scholars have given ample attention to the need for more comprehensive and systemic designs to properly study the effects of complex interventions in school settings. Recommendations of Donald Campbell (1981) and Lee Cronbach and associates (1980), in particular, speak to the utility in *mixing* various research methods, and in employing "before" and "after" studies that build on one another *over time* to address questions of program impact. Such writings point to a clear need for researchers to judiciously combine comparative, qualitative or descriptive research methods to properly answer questions on how a novel program might work, what it looks like in operation in early and later stages of implementation, the conditions under which it influences particular outcome measures, and the likelihood that it will work in the same way with other students, across settings and over time.

Federal Recommendations for Schools to Use Supplemental and Extended Day Services

Supplemental programs. The U.S. has had a long history of providing supplementary education via schools, community organizations, churches, for-profit education providers and other agencies to students in all achievement and socio-economic brackets. However, the press for schools to use supplemental instruction as a strategy to benefit economically disadvantaged, low-achieving minority students heightened in the past decade of standards-based education reforms in the U.S. The No Child Left Behind Act of 2001 (PL 107–110) expanded the range of service options for parents whose children attended Title 1 schools that were flagged as needing improvement. NCLB defines supplemental educational services as tutoring and "research-based" academic enrichment programs that supplement, but do not replace, instruction provided by schools during the school day.

Among the choices offered under the law, children from low-income families enrolled in schools not making adequate yearly progress (AYP) for two consecutive years are eligible to receive supplemental educational services, including tutoring, remediation, and other academic instruction. Under the NCLB Act, supplemental education service provision is to be overseen by states. To facilitate state-level implementation in 2002–03, the U.S. Department of Education (DOE) issued non-regulatory guidelines to assist schools and school districts in selecting and monitoring supplemental service providers as well as in gathering evidence of program/provider effectiveness (www.ed.gov/policy/elsec/guid/suppsvcguid.doc)

NCLB's broader strategy for fostering school improvement and accountability calls for under-performing schools to offer "supplemental educational services" for students failing to meet standards on external accountability tests administered by states. Approved programs, funded through Title I and provided to students in schools that do not make AYP for three consecutive years, are required to show increases in student achievement levels, with schools attaining correspondingly higher performance standards set according to state criteria (P.L 107–110, 115 Stat. 1425, 2002).

A recent federal report released data on the implementation status of supplemental instruction programs by states under the NCLB Act (Anderson & Weiner, 2004). The study used a telephone survey method and found that generally, states were complying with DOE guidelines in selecting supplemental providers; districts and schools were making strides towards implementation;

but little evidence was found of any systematic efforts to monitor provider effectiveness at either the state, district, or school level.

Other than the NCLB, a spotlight on supplemental education is also found in recent recommendations of the National Task Force on Minority High Achievement convened by The College Board (1999). The Task Force's report carries a clear message that a viable means for poorly-achieving minority students to improve their academic achievement is by employing after-school supplemental strategies that have proven success with "educationally sophisticated or savvy" parents and student groups (p.18). Schools have several options when it comes to commercially-distributed supplemental instruction products, including the one investigated in the present study.

Extended day programs. An associated reform initiative prompted by NCLB is extended-day schooling. Extended-day programs generally take the form of schools adding an hour or two of supervised schooling during which all or selected groups of students are provided with after-school care and/or tutoring services in academic subjects. Based on the Schools and Staffing Survey data collection conducted by the National Center for Educational Statistics between 1990–94, DeAngelis and Rossi (1997) reported that extended-day programs have increased greatly in U.S. elementary schools over time and are now serving greater numbers of minority and high-poverty students. However, such programs were fewer in number in rural than in urban schools, and among private institutions, their availability is greater in Catholic schools.

Not all extended day programs provide supplemental instruction, devoting time instead to supervised extra-curricular activities. There is some descriptive evidence from a number of large efforts, including the Big Brothers and Big Sisters of America mentoring program, that show improved academic achievement on standardized tests such as the *Stanford Achievement Tests* (9th Edition), better school attendance, and improved psychological and behavioral outcomes for at-risk youth, such as reduced gang-related behaviors, violence, or drug use (University of California at Irvine, 2001; Aguirre International, 2000; Huang et al., 2000; Grossman et al, 2000). To achieve success on academic outcomes, Owens and Vallercamp (2003) isolated the following five major factors that extended day programs should embody: addressing identified needs within a school; building on a shared vision among the school and larger community; fostering staff ownership; having ties to state curriculum standards; and measuring and sharing results across the community.

Available evidence on the effectiveness of various supplemental instruction programs and the best models for their delivery in urban schools and large city school systems is still somewhat sparse. Few rigorous evaluations exist, according to a recent report of a national Task Force on promotion of minority achievement (The College Board, 1999). The success of supplemental programs, according to Cohen (2003), is predicated on several factors, such as a strong parent, tutor, and teacher connection; experienced providers and developers; proven methods of instruction; customized instruction; measurable results based on time on task; and positive learning environments. Although choices exist, available information on program efficacy is still mostly anecdotal, with formally-gathered research evidence limited on effects of various supplemental programs in different populations. One large-scale federally-supported study, discussed next, is an exception.

The 21st Century Community Learning Centers (21st CCLC) Evaluation

Evaluation design and findings. To raise achievement levels in disadvantaged and struggling students, the Elementary and Secondary Education Act supported supplemental center-based programs in over 360 rural and inner city schools in 34 states in 1998. Labeled as the 21st Century Community Learning Centers (21st CCLC) initiative, this program of supplemental education was

reauthorized under the auspices of NCLB in 2002, with an additional one billion dollars. In 2003, DOE released its first year findings from the 21st CCLC national evaluation examining program characteristics and outcomes (Mathematica Policy Research, Inc. & Decision Information Resources, Inc., 2003). This study, although labeled as "first year findings" was conducted after the initiative received three years of funding.

The national evaluation of the 21st CCLC utilized a randomized experimental design to ascertain effects in some if not all centers (Mathematica Policy Research, Inc. & Decision Information Resources, Inc., 2003). The evaluation's design incorporated separate studies with middle and elementary school students. The elementary study used random assignment of students to treatment and control groups in 14 school districts with 34 centers; the first year study focused on data from 7 of the districts grantees that could implement the experimental design; data from 1000 randomly assigned students were analyzed (Mathematica Policy Research, Inc. & Decision Information Resources, Inc., 2003, p.13). The middle school study used matched samples of students in treatment and comparison groups; it focused on 62 centers in 34 school districts. Evaluators collected baseline and follow-up data on 4400 middle school students from 32 of the district grantees. In addition, 2–4 day site visits were conducted to gather supporting data on program profiles in both elementary and middle school studies. Outcomes were measured on students' perceptions of safety, attendance, test scores and grades in academic subjects, and teacher satisfaction with homework or class work completion.

Implementation findings showed that programs were staffed by school-day teachers on additional pay and offered 4–5 days a week but lacked in academic content. Markedly, programs posted low student attendance rates (an average of 2 days per week) and were limited by inadequate plans for sustainability, according to the authors. Little or no differences were found between the treatment and comparison students on any of the outcomes at both elementary and middle school levels at the end of the third year of implementation.

The 21st CLCC evaluation design and interpretive constraints with results. The authors of the 21st CCLC report describe their study as "one of the few" that are consistent with NCLB criteria for scientific rigor because of their use of randomized trials (Mathematica Policy Research, Inc. & Decision Information Resources, Inc., 2003, p.xiv). At the same time, they admit to many shortcomings of even their elementary-level investigation where they reported the use of RCTs. Among others, their reported concerns surround the lack of sample representativeness, limited generalizability of results, cohort differences by year over the period of implementation, and student similarities/dissimilarities stemming from nestedness in school-based centers across multiple districts (Mathematica Policy Research, Inc. & Decision Information Resources, Inc., 2003, p.13).

Other methodologists or stakeholders could raise additional questions. First, because selection of control students was dependent on surplus enrollments at funded centers--a logistical barrier—the researchers could only employ RCT at the elementary level. Second, there were no significant effects after 3 years of program implementation nationally, but interpretations of the effects were difficult to make based on the limited information collected on ongoing program inputs, processes, local environmental dynamics and variables. Finally, the effort sought definitive information on effects without any built-in attempt at providing formative feedback to strengthen program delivery as the centers became established. Thus, while the scope of the information targeted by the study as a whole was huge and the costs of a multi-site, multi-year national evaluation enormous, the evidence obtained within and across sites was superficial at best—constrained by the scale of the effort.

Too much faith had been placed on the "magic" of randomization in the 21st CLCC elementary level investigation. There was no empirical verification of sample equivalence over time nor of contextual irregularities or variability in treatment and control conditions within and across

sites over three years of implementation. Multiple cohorts appeared to be mixed up in that study. Data on program characteristics were gathered *post-hoc* through brief site visits. No first-hand documentation or data existed on qualitative differences in various models of program delivery as they emerged in actual school environments; no direct links could be made between particular program characteristics and particular outcomes. Some centers may have been more effective than others, and some may have had better attendance than others, but such differences were clouded in the results.

While the researchers did a good job of documenting several limitations in their procedures; randomization as a procedure got severely compromised in the field application and did not help them in their cause to gather high quality evidence on program effectiveness. Besides the documentation that participation rates had been uneven and low—other factors that may have explained the disappointing results remained in a "black box".

Almost immediately after the release of the study, federal funding for the 21st CCLC was cut by 40%. The drastic action catalyzed interest in developing a stronger "research and evaluation agenda" that allows for continuous improvement of similar innovations as well as accountability to funders (Harvard Family Research Project, 2003, p. 1).

Essential Elements of the ETMM Approach

While RCTs (like the one described) often target multiple sites across the nation to obtain statistically desirable sample sizes for hypothesis testing, they give minimal attention to program processes and environmental factors in their design. ETMM designs, in contrast, are guided by a program's theory of action and mix research methods. They complement field experiments with ongoing observations, interviews or survey research to better gauge how relevant variables might affect outcomes. The aim of such designs is to document relevant facets of a program as it operates in its natural environment, as systemically and comprehensively as resources will allow. The research plan in ETMM designs deliberately targets a significant portion of the life of an intervention for study, incorporating two self-contained phases of work: an exploratory, formative investigation, followed by a confirmatory, summative investigation. The formative phase is used to provide feedback to program participants to shape program delivery, to better study the treatment, control conditions and the environment, as well as to improve the research design as more is learned empirically about the larger context in which a new program operates. The summative phase incorporates more formal experimentation. Together, the two phases in an ETMM design are intended to yield a comprehensive body of evidence that permit researchers to make sound determinations of impact with knowledge of conditions under which the effects were manifested (see Chatterji, 2004, 2005, for design principles).

A Demonstration of the ETMM Approach with a Supplemental Program Evaluation

The present ETMM application was constrained by limited resources and is thus a less than "ideal" implementation example. However, it still yielded a corpus of evidence that facilitated a more holistic appraisal of likely effects of the supplemental program under similar conditions than would a traditional RCT. The research involved a year-long study and combined a matched-groups, quasi-experiment with classroom observations and surveys. This design was implemented in two successive phases of research. A 14-week formative phase explored the program and its

environment in depth and was aimed towards providing feedback to developers, program personnel and school staff so as to stabilize treatment delivery and improve fidelity. That was followed by a 16-week summative study of short-term and very early impacts, where findings of the first phase were used to tighten the data-gathering and analytic design in specific ways. Details of the context, methodology and findings follow.

Context of the Evaluation Study

The present study was conducted during the 2001–02 academic year and was a pilot of the program in New York City schools. The *treatment program* was delivered as a component of the extended time schooling initiative already under way at the school site. The school, located at Harlem, had been marked as a school under review by the city board of education in the previous year. The school administration hoped to improve student performance on state and city tests in all grades from Pre-K through 5. The program was one of several reform initiatives concurrently being implemented by the school to achieve this objective.

The research was initiated in response to a request from the program developer. The broader stakeholder group included the principal, teachers, students and parents of the school, all of whom were engaged in the deliveryor utilization of evaluation results to some degree during the pilot year, along with the provider. The primary goal of the research was to comprehensively examine how well the program performed in a New York public school environment. The more typical setting for the *treatment program* consisted of after-school community centers, where participating children were from the middle to high socioeconomic brackets, and active parent volunteers ran the program. For the first time, the program was being tested with ethnic minorities in New York City, all of whom were enrolled in the free or reduced lunch program at the Harlem public school (i.e., in the low socioeconomic bracket). Most were struggling in reading, mathematics, or both subjects.

Treatment Program Characteristics

The program (referred to as the *treatment program* hereafter) is described by the developers as being among the world's largest providers of supplemental education materials. The method emphasizes computation in mathematics and basic reading skills, the development of speed and accuracy skills through practice and repetition, independent learning, and self-paced mastery of graduated materials in basic mathematics and reading. The program incorporates some characteristics associated with potentially successful supplemental programs mentioned by Cohen (2003), in that it attempts to involve both parents and teachers in school-based delivery models, allocates blocks of work time for students, and matches student levels to materials through initial placement testing. Others have noted that the program aims to make basic skills, such as computation, automatic by promoting over-learning shaped by feedback, and uses timed conditions that mimic conditions of standardized testing (Weischadle, 2002).

The supplemental curriculum in reading and mathematics was delivered in 20-minute work blocks in each subject, three days per week, during the extended hour of the school day in treatment classrooms of the school site. That is, it was selectively delivered as a component of the extended day schooling initiative already in operation at the school, in particular treatment classes. Teachers in treatment classes volunteered to participate during the pilot year following schoolwide training and orientation activities that occurred in the preceding summer. In comparison classes, by contrast,

students did not receive the supplemental program during the extended hour of schooling or at any other time.

The supplemental curriculum consisted of sequenced sets of multi-item worksheets (referred to as *assignments* by the developers), founded on the philosophy of its developers. To start, children were given placement tests and started by the developers at levels that matched their ability levels on specific subjects. Children were expected to progress at individualized paces through the leveled assignments on their own, with minimal guidance from teachers/ facilitators. They followed a set daily routine, where they were expected complete assignments under timed conditions. Before each session, they reviewed their homework, re-did or corrected missed problems from the previous session, and moved on to the new worksheet assigned. Per program theory—or the underlying assumptions on which the program was built--expected outcomes were higher levels of reading and mathematics achievement, self-efficacy as evidenced in their self-reports and confidence in attempting more tasks/items, better completion times, and independent work habits. Nine classrooms, ranging from Pre-K through Grade 5 and including one, mixed-grade special education class, participated in the program during the year of the study.

Treatment Program's Underlying Theory

The design of the study began with an analysis of the supplemental program's theory of action or the set of explicit or implicit assumptions that suggested how the desired outcomes would be affected by variables in their context and the program inputs and processes (after Bickman, 2000). The major components of the supplemental program's theory were extracted by the research team based on a qualitative review of the program materials, videos, documentation supplied, and ongoing consultations with staff of the curriculum corporation. These findings were organized under Program Inputs (resources and services allocated to set up and run the program at the site), Program Processes (activities that were expected to occur as a result of the inputs), and student and program outcomes that were expected to ensue.

The logic model (Figure 1) depicting the *treatment program*'s theory shows that the supplemental program aimed for the same achievement outcomes as the regular school-day's programs in reading and mathematics. Critical context variables to consider in the design, delivery and analysis of the supplemental program were student characteristics and the urban location of the school, along with its status as a school under review in the city system. As shown, multiple school-wide initiatives were concurrently in effect to raise student achievement at the school when the study commenced. The key ones included smaller class sizes (a structural/organizational intervention), the regular-day reading (*Success for All*) and mathematics curriculum (curriculum/instruction interventions), school-wide parent involvement incentives and an after-school snack program for children during the extended hour of school (student services/support interventions). In terms of inputs, the additional total cost of the *treatment program* in a given subject area per child was reported to be approximately $300 in a 9-month school year. More specifically, inputs during the after-school sessions for children receiving supplemental education could be classified under five major headings.

Placement testing. To begin the program, students were placed at a level in which they were most likely to succeed in a particular subject area supplemental curriculum. Placement tests were administered to each participating student and scored by the developer's staff to achieve this purpose.

Materials. The program in each subject area consisted of assignments focusing on leveled basic skills. These assignments were kept in storage shelves provided by the developer, and housed

in a resource room provided by the school. Additional supplies included posters, number games, and other materials intended for skill-building relevant to the supplemental curriculum. Periodic achievement tests were administered to students focusing on blocks of completed worksheet skills. Student performance reports, prepared by the corporation, were supplied back to teachers, parents, and students following achievement testing. Rewards and recognition systems were implemented to keep students motivated.

School Context Variables Inputs and Processes Outcomes

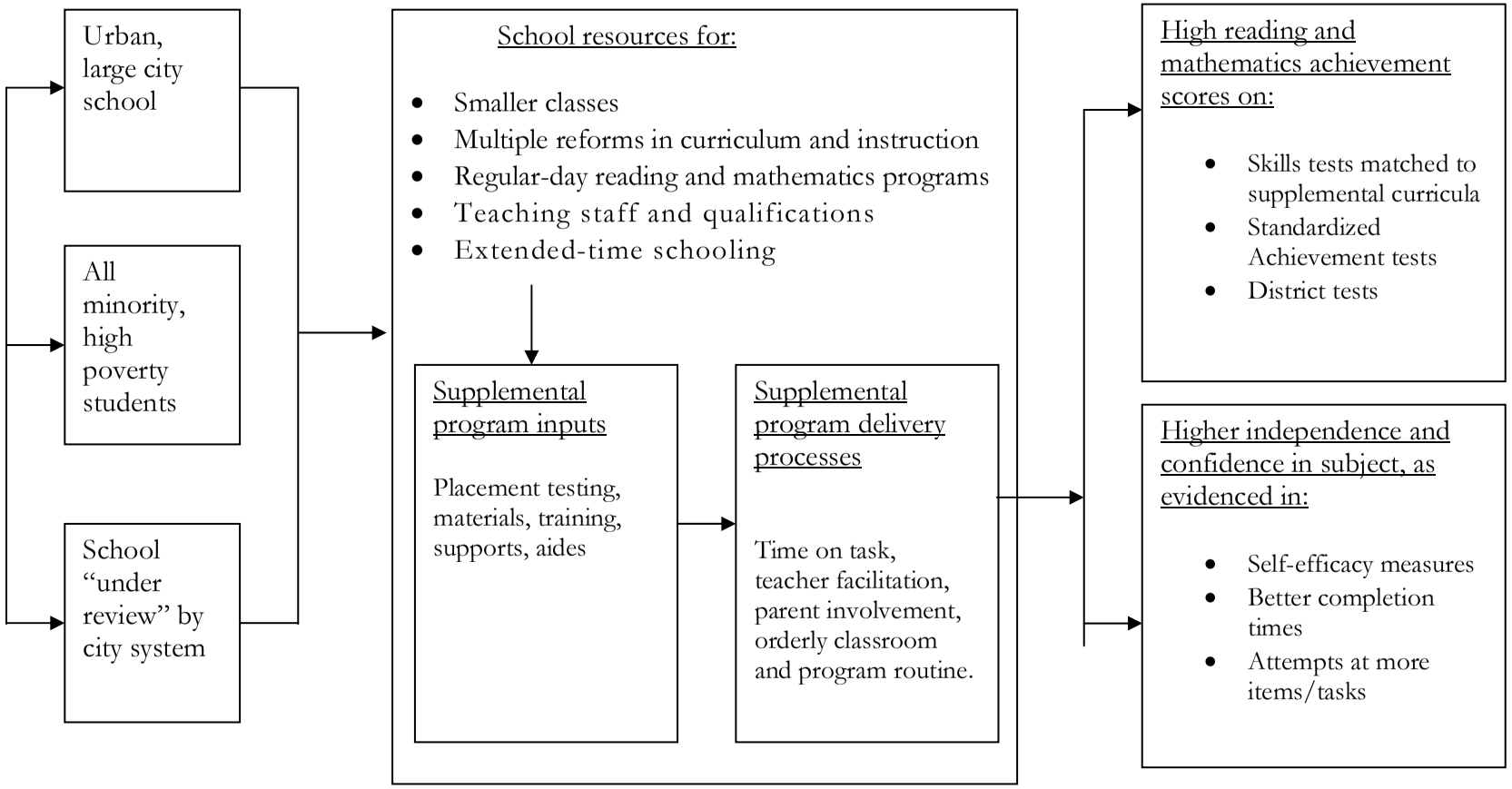


Figure 1. After-School Supplemental Program Theory Model

Training and support services. Developers provided school administrators and teachers in all participating classrooms with training and materials before the program began. The corporation's staff provided ongoing assistance to teachers and helped with program organization and delivery throughout the first semester and for much of the second.

Aides/assistants: The corporation also provided aides/assistants to assist with the daily grading of assignments and management of materials in *treatment program* classrooms.

As evident in Figure 1, several *treatment program* processes were expected to occur as a result of the inputs. Among the critical ones were the following.

Student time-on-task. For participating classrooms, the after-school hour was broken down into 20 minute work blocks in reading and mathematics, respectively. Children were expected to follow a structured routine to complete assignments for at least this period of time on days with supplemental instruction.

Teacher-facilitated delivery. Following the diagnostic testing, individual classroom teachers were responsible for program delivery based on the prescribed program philosophy and daily regimen. Once trained, teachers were expected to allow individual children to complete each day's assignments as independently as possible. Although not expected to score student assignments, teachers were expected to provide the feedback and coaching needed to help individual children begin their work each day, or correct mistakes from the previous day's work. Teachers were also expected to manage students' classroom behaviors during the supplemental hour, including keeping children occupied once worksheet activities were completed for the day.

Parent involvement. The program aimed to actively involve parents in their children's learning. To that end, the corporation's staff held parent orientation meetings, sent homework sheets home with particular children, and prepared student reports for parents.

Orderly classroom environment. Videos of ideal classrooms depicted an environment that was quiet, organized, and orderly, with children needing very little one-on-one guidance. When the program operated according to guidelines, teachers/facilitators were minimally involved, and students progressed from level to level guided by their own high motivation and engagement levels. The classrooms were expected to be distraction-free and conducive to independent learning.

Other *treatment program* assumptions were implicit. The after-school curriculum was intended as a supplement to the regular curricula in reading and mathematics, emphasizing state content standards. Thus, there was an implicit assumption that the embedded skills would be aligned with and complement those typically covered by teachers in Pre-K through Grade 5 classrooms during the regular school day. The regular-day curriculum was also expected to affect children in *treatment program* and *comparison* classes uniformly. Once inputs were allocated, it was assumed that there would be consistent levels of support and buy-in from teachers, school leaders, parents, and students, so that the program ran smoothly, as designed. Because of the emphasis on parent involvement, more parents were expected to be involved in their children's education in the supplemental program classrooms than in classrooms without these services.

Evaluation Questions

Given the program's theory of action, questions that guided the design and data gathering procedures were classified under four headings: treatment fidelity (both formative and summative phases), teacher perceptions and buy-in (both phases), initial process-outcome relations and moderator effects (formative phase only), and early treatment impact (summative phase only). Questions are listed below.

Treatment fidelity. To what extent were inputs and processes observed during the pilot year, consistent with theory in *treatment* classrooms? Were program inputs and processes observed in *treatment* classrooms changing over time in directions expected per program theory?

Teacher perceptions/ buy-in. Did participating teachers report satisfaction with the program products and services in the early and later phases of program implementation?

Initial process-outcome relations and moderator effects. Did the treatment yield better achievement outcomes for comparable groups of children in the formative phase? Did children's achievement vary in *treatment* versus *comparison* classrooms where teacher perceptions on selected environmental variables varied (i.e., were high versus low)? These variables included perceptions of alignment of the supplemental program with the regular-day curriculum, observed parent involvement levels, and observed levels of student independence.

Short-term treatment impact. Controlling for mid-year achievement, were there short-term effects of the supplemental program in reading and mathematics on key outcomes in comparable *treatment* versus *comparison* children?

Methods

Because the supplemental program was individually adapted, students at a given grade level were permitted to start at different points and move at varying paces through the after-school curriculum. To target both the primary and intermediate groups, parallel forms of multi-level achievement tests were designed in each subject area to serve as outcome measures. These tests were expected to be more sensitive to early effects of supplemental services. Methods for observing and recording all input, process and outcome variables described next were the same in both phases of the research .

Formative Phase—The "Before" Study

The formative study of the program began soon after the summer teacher orientation. It yielded documentation of the extent to which the observed program processes, inputs, and outcomes were consistent with the program's underlying theory and philosophy in the very early life of the program (semester 1). Process data were gathered using classroom observations and teacher surveys, along with outcome data on multi-level reading and mathematics tests focusing on skills reinforced through the *treatment program*. Matched samples of treatment and comparison group students by primary (Grades Pre-K-1) and intermediate level (Grades 4–5) were identified at the start of the school year. All children were first-time enrollees at the particular grades and not in special education. A Grade 3 class with retained students and a special education class did not have matches by grade and were treated separately to improve internal validity of the comparative design (descriptive data were collected for them). In the comparative design, thus, the primary and intermediate samples were essentially independent samples matched by grade; demographic equivalence of the within-grade samples was examined at the start, but could not be sustained due to student mobility (detailed next).

Descriptive analysis of the qualitative and teacher survey data were complemented with two-way ANOVAs that examined early process-outcome relationships by grade, with appropriate moderators as independent factors (e.g., effects of high and low levels of teacher-perceived curriculum alignment with the supplemental program by treatment versus comparison group). The outcome analyses used grade-free multilevel skills tests as the main achievement outcome measures

in reading and mathematics. The multi-factor ANOVAs helped examine and as necessary, rule out effects of extraneous environmental factors on student achievement and select an optimal statistical design in the summative phase. In addition to informal exchanges that occurred regularly between the teachers, researchers, the developers and school personnel, results of the formative study were formally fed back to program developers, sponsors, and on-site participants as program implementation continued in mid year.

Summative Phase—The "After" Study

At the request of the sponsor, the summative phase of the evaluation was implemented during the last 16 weeks of the school year as program implementation continued. It was also guided by the program theory model. Data collection continued with classroom observations and surveys to document changes on program inputs and processes over time in matched classrooms by grade. Using the end-of-first semester scores on different subject area tests as the covariate, ANCOVA and effect size comparisons were now used to draw conclusions on early program effects in the previously identified treatment and comparison students within independent, primary (Grades Pre-K through 1) and intermediate level (Grades 4–5) sub-samples. Student mobility and attrition rates that the school and researchers were unable to control, reduced sample sizes in the summative phase. Corrective actions included the use of the mid-year covariate to equalize pre-existing domain-specific student differences in the summative analyses.

The data were checked to see if homogeneity of regression assumptions for conducting ANCOVA were met (i.e., there was no interaction between the covariate and treatment conditions). Independent factors in the first analysis were treatment versus comparison conditions. Dependent or outcome measures were reading and mathematics scores on the multi-level tests. Effect sizes were computed using Glass' formula to understand the direction and magnitude of initial effects. Additional analyses compared means descriptively on other outcomes in treatment/comparison groups.

Changes in Comparative Research Design

The present ETMM application incorporated a comparative design that has been characterized as a *quasi-* rather than a *true*-experiment. While students in the school were "randomly assigned" to teachers in the beginning of the school year because of an administrative policy of heterogeneous grouping, 9 of the teachers (classrooms) volunteered to participate in the treatment program across grade levels—this resulted in uncontrolled conditions with respect to teacher equivalence in treatment and control conditions.

In matched classes by grade, however, equivalence of students from *treatment* and *comparison* conditions was attempted and periodically checked on four background characteristics: ethnicity, gender, membership in free lunch program, and native language spoken at home (Limited English Proficiency status). Initial equivalence was established within grades.

To obtain higher sample sizes by level, a decision was made to separately study primary (PreK-1) and intermediate (Grade 4–5) samples using students from combined grades at each level. Grade-level breakdowns were examined descriptively prior to initiation of the formative study, and grade was used as a control variable in later statistical analyses. Because the primary matching variable was grade level, the samples were treated as independent samples in statistical comparisons

and hypothesis tests, with covariates included in the summative analyses. Due to small numbers, nestedness of students in classrooms was not taken into account in the analysis.

Subject Characteristics

Table 1 shows *treatment* group statistics on mean number of assignments completed as an index of program exposure. Tables 2 and 3 show the characteristics and numbers of students in samples at the point of commencement of the formative study, and in mid year before the summative phase began.

During the course of the investigations, attrition due to student mobility, inadequate exposure to the *treatment* due to irregular attendance, or missing data on critical outcome variables resulted in changes in sample composition and fewer cases for particular summative analyses. These changes to sample size reduced power of the statistical tests in the summative phase, but *did not* markedly alter the comparability or representativeness of the original matched samples on background characteristics deemed relevant for the investigation (this was checked, and proportions were comparable in different ethnic and gender groups). Regardless, because of sample attrition, summative analyses incorporated a covariate to adjust for mid-year differences in academic skills in both subject areas and used the adjusted Sums of Squares (Type III) for calculation of variances because of unequal Ns in cells.

Table 1
Mean Treatment Exposure by Grade , Subject Area and Level: Number of Students

Grade and Subject	Mean # of Assignments	*SD*	*N*
Pre-K			
Reading	510.77	155.64	13
Math	703.77	170.20	13
Kindergarten			
Reading	654.67	208.52	15
Math	682.67	245.40	15
Grade 1			
Reading	706.33	210.32	15
Math	931.47	165.11	15
Grade 4			
Reading	700.00	167.52	20
Math	673.50	182.56	20
Grade 5			
Reading	706.50	177.56	20
Math	733.75	161.71	20

Table 2
Demographic Equivalence in Initial Treatment and Comparison Samples by Level

Demographic Variable (Level)	Treatment Group	Comparison Group	%	*N*
Gender (Primary)				
Male	21	21	45	
Female	26	26	56	94
Gender (Intermediate)				
Male	15	15	48	
Female	16	16	52	62
Ethnicity (Primary)				
Black	31	31	66	
Hispanic	8	8	17	
Other/Unknown	8	8	17	94
Ethnicity (Intermediate)				
Black	28	28	90	
Hispanic	3	3	10	62
Free/reduced lunch (Primary)	47	47	100	94
Free/reduced lunch (Intermediate)	31	31	100	62
English speakers (Primary)				
Non-native English speakers	7	7	15	94
Native English speakers	40	40	85	
English speakers (Intermediate)				
Non-native English speakers	3	3	10	62
Native English speakers	28	28	90	

Table 3
Sample Sizes by Level in Mid-Year prior to Summative Study

Level (Outcome Measure)	Treatment	Comparison	Total N
Primary (Reading)	35	33	68
Primary (Math)	35	33	68
Intermediate (Reading)	30	31	61
Intermediate (Math)	29	30	59

Breakdowns by grade available on request.

Sampled Observation Notes	**Codes Consistent with Program Theory**
Pre-K classroom: Children seem to know what to do. Children in groups of 5–6 at table with adult–aide or teacher.	[Ss following program protocol]
Each focused on worksheet. Engaged in worksheet. Aide guiding student to write– "down up, down up..."	[Ss and aide following program protocol]
Grade 4 classroom: One girl has finished her worksheet, She says to observer " XXX has helped me in math."	[aide assisting Ss; positive environment]
She walks to desk, checking her sheet. She discovers she has missed 3 items. She returns and begins to do them	[Ss following program protocol; positive comment on program]
Sampled Observation Notes	**Codes Inconsistent with Program Theory**
Grade 1 classroom: (Developer) giving directions for XXX routine to students..."be quiet, get your packet, get ready for XXX. But no one seems to pay attention to him, except for a few kids. They are extremely noisy....	[Ss loud; developer managing Ss' behavior]
Grade 3 classroom: Only 5 students in class; 4 of whom are on task.	[Supplemental program attendance low]
Teacher working hard–trying to keep them seated.	[T managing Ss' behavior]
R says–"Shouldn't the sheets be matched to their levels of comfort?" Teacher responds– "Yes, but we moved them up faster and they are discouraged"	[Ts not following program protocol]

Figure 2. Classroom Observation Records and Line-by-line Coding Procedures

Data Sources, Measures and Data Collection

Details on the development and validation procedures for three newly developed instruments are given under particular sub-headings. The appendix provides additional details on assessment specifications and items with early validity and reliability data.

Classroom observations of program inputs and processes. Narrative running records of *treatment* classroom activities were sampled during the supplemental hour by observers at both

primary and intermediate levels. For the formative study, a total of 20 such observations were conducted for 30 minute periods each and distributed equally in intermediate and primary classrooms. Likewise, in the summative phase, 11 observations were conducted (5 were in primary classrooms, 5 in intermediate classrooms, and 1 in the grade 3 class). The text data were coded line by line, using classical content analysis procedures (Ryan & Benard, 2001) and codes were clustered under general themes.

A sample of coded observation data is shown in Figure 2 and illustrates how codes extracted from each line of text data were classified under broader themes to evaluate their consistency with expectations given by the program theory model in Figure 1 (results reported in Table 4). To examine changes over time, the proportions and rank-order of counted codes by theme category were compared in the first and second semesters of program implementation, the formative and summative phases of the research.

Teacher self-report surveys in participating and comparison classrooms. In both semesters, *treatment* teachers were asked to rate the quality of different aspects of the supplemental program. At the end of the each semester, teachers in both participating and non-participating classrooms matched by grade level (N=20, 8 in paired classrooms by grade plus others) were also asked to respond to items tapping three key moderator variables: perceived alignment of the supplemental program with the regular curriculum in reading and mathematics, perceived parent involvement levels in their classes, and perceived levels of student independence.

To check for their perceptions on the degree of regular curriculum alignment with the supplemental program objectives, skills were extracted through a content analysis of the supplemental materials, and presented to treatment and comparison teachers in the survey (the complete instrument appears in the appendix). Item responses and means on survey indices were compared descriptively in *treatment* and *comparison* classes in both semesters to obtain a sense of the differences on contextual variables under the two conditions.

Table 4 shows sample items from each sub-domain of the survey. As is evident, Cronbach's alpha reliability estimates were found to range from .73-.89 (greater than the acceptability criterion set for .70) on all teacher survey indices.

Student outcome measures. Student achievement scores, time taken, and number of items attempted, on the specially designed multi-level skills tests in reading and mathematics were the outcome measures used to evaluate short-term effects of the program. The domain for each test was ordered, and represented by progressively complex groups of skills, starting at the beginning of pre-kindergarten levels and going to a few levels beyond the maximum achievement expected at the highest grade. Test specifications, shown with sample items in the appendix, were developed with the involvement of staff from the curriculum corporation. Items matched to each skill area were then selected from the existing pool of published curriculum materials. Because of the volume of assignments and items published, prior exposure to items was not considered to be a major threat to student performance measures obtained.

Two parallel forms of each multi-level test were prepared at each level and subject area, for separate use in the formative and summative phases of the study. Split-half reliability of the forms, based on a separate pilot study with a center-based sample, ranged from .67 to .72 in the primary group, and .78 to .82 in the intermediate group in reading and mathematics, respectively. Convergent validity coefficients with supplemental program exposure, items attempted, and speed of completion were moderate to high and consistent with theoretical expectations (reported in the appendix).

Table 4
Teacher Perceptions Survey: Results in Treatment and Comparison Classes

Dimension *Sample item*	Total Items	Phase of Study	Treatment Mean	Treatment SD	Control Mean	Control SD	Cronbach's alpha
Curriculum alignment, reading *To what extent is the following skill/area addressed in your regular curriculum: Identifying main ideas?*	6	Formative Summative	15.17 15.10	3.13 3.14	15.11 15.51	2.47 2.92	.89
Curriculum alignment, math *To what extent is the following skill/area addressed in your regular curriculum: Sequencing numbers?*	11	Formative Summative	14.50 24.11	3.73 6.08	14.00 22.61	4.27 5.44	.85
Parent involvement *(In your class) To what extent are your students' parents/guardians involved this year in: Helping students with homework?*	3	Formative Summative	6.54 6.42	1.94 1.81	6.89 7.10	2.13 1.86	.74
Student Independence and Goal-directedness *(In your class) To what extent are your students' showing independent, self-directed behaviors in Mathematics?..In Reading?*	4	Formative Summative	7.67 8.72	1.97 1.22	8.56 9.51	1.88 1.61	.74

Formative Phase N (teachers)=15; Summative Phase N (teachers)=19

Evidence of content-based validity (match of tests' content with teachers' regular-day curricula) of the skills sampled on the multilevel tests in reading and mathematics was obtained by semester through the teacher survey, and is shown in the appendix. As is evident, teachers in both treatment and comparison classrooms saw greater alignment of the reading skills with their regular curriculum than with mathematics skills; however, as the school year progressed, more of the mathematics skills were covered by teachers in both conditions, improving content validity by the end of the summative phase (see increase in composite score mean on curriculum alignment in Table 3).

Test administration conditions were un-timed. Each child started at several levels lower than their assessed ability level and was asked to go as far as he or she could. Starting and ending times were recorded. Scoring was standardized with the help of a key, and included partial credit scoring on a few items. Scorers were formally trained in a practice session until they were found to agree on their scoring decisions. Levels of scorer agreement in scoring of particular items was found to exceed 70% with practice tests.

Other outcome measures. Two self-efficacy scales (see the appendix), focusing on reading and mathematics respectively, were developed and validated for use in the summative phase of the study. Based on indicators drawn from the theoretical literature on self-efficacy, these instruments included 13–16 self-report items with 3 point Likert scales. A typical item asked, *Can you do the math problems your teacher gives you?* The primary level instruments were designed as interview-based assessments, while at the intermediate level the same instruments were administered as teacher-guided paper and pencil questionnaires. The intermediate level self-efficacy scales were content-validated against theoretically derived indicators by external experts and the research team. The scales showed adequate Cronbach's alpha reliability (.74 in math and .77 in reading). The primary-level instrument was tested during the formative investigation but not used in the summative study due to unacceptable reliability.

Finally, scaled scores from the state and city standardized achievement test, CTB-4, were also used as additional measures of achievement outcomes in the second phase at the intermediate level. For primary children, teacher ratings from the *Early Childhood Language Arts Scale* locally-developed in the New York City system were used to compare *treatment* and *comparison* students.

Program Fidelity in Formative and Summative Phases: Changes in Treatment Definitions

In the formative study, potency of the treatment was operationally defined based on the number of after-school sessions attended by *treatment* children, with data collected on number of worksheets completed to supplement that information. However, site observations during the formative phase revealed that not all students attended the after-school supplemental sessions regularly. Further, they were often pulled out early by their parents who took the assignments home for completion. The school principal added Saturday sessions to the extended hours on school days. The providers allowed this to happen, as it fit their program theory calling for greater parent involvement and task engagement.

A change was thus made to the summative study to improve validity of the design. An *a priori* decision was made in consultation with the providers and school stakeholders to set a cut-off for student exposure to *treatment* at a minimum of 100 assignments in a subject area and to a minimum of 200 assignments over two semesters. Thus, the "treatment condition" was now operationally defined in a broader way based on task completion both in and out of the after-school classroom environment. This resulted in a small change in the composition of the original samples at the primary and intermediate levels in the summative phase (fewer than 10 students were excluded, and most of these had moved away from the school). Instead of imposing a standardized model that could not be sustained in real school environments, this alternate program model was collaboratively considered a more realistic operational definition of the supplemental program.

As indicated earlier, to enhance internal validity of the quasi-experiment, key extraneous variables identified in the environment were examined statistically and ruled out as possible threats before the comparative summative study was undertaken. Grade-retained students without similar matches received year-long supplemental services in Grades 2 and 3 (N=11 in each). Likewise, a mixed-grade special education class without matching pairs of children were in the supplemental

program (N=7). These students were studied as separate samples using one-group, pre-test to post-test change designs. The analyses were treated as descriptive, because of the lack of matching comparison children and small sample sizes. The summative study of preliminary program effects thus focused on a primary sample (Grades PreK-1) and an intermediate sample (Grade 4–5) and used a comparative design, matched by grade level, and controlling for mid-year achievement on multi-level math and reading tests as the covariates.

Results

Extent of Treatment Fidelity: Classroom Observations

At the end of the formative phase, classroom observation results were mixed (see the left panel of Table 5 showing frequencies). However, classroom processes changed in positive directions by the end of the year (Table 5, right-hand panel showing frequencies). The percentages in Table 5 refer to proportions of the total coded text data in different thematic categories by semester. Examples of text segments under each theme are provided as quotes in the extreme left-hand column. Themes have been logically grouped under broader "input" and "process" categories. Results from the formative phase in Table 5 can be compared on common thematic categories with results of the summative phase using rank-orders, rather than the absolute frequencies, as the number of observations lessened by about 1/3 in the second semester. The summary results reflect activities documented in classrooms sampled by semester; primary and intermediate level data are combined in the table.

Table 5 (left) shows that program inputs were largely consistent with theory in the formative phase–with both the developers and the school principal jointly investing considerable resources. The principal and corporation staff were documented to be highly involved with program delivery. Most teachers and aides were involved in classroom practices that were consistent with the program theory, although some of their actions were directed towards arresting student misbehaviors. Classroom processes were uneven, however, particularly in intermediate classrooms (grades 4–5, not isolated in the table). In all, there were 240 (41%) coded occurrences of student unruliness and 61 (11%) associated classroom management behaviors. Such observations were classified as inconsistent with the theoretical expectations of a smoothly operating and quiet classroom. Among other inconsistent findings, parents were often observed pulling their children out during the supplemental hour and teachers tended to let them take assignments home.

At the end of the summative phase (right hand panel of Table 5), observational records showed patterns suggesting that the program was being implemented in a manner that complied *more* with the major program guidelines. Notably, behaviors of students and teachers, at both primary and intermediate levels, were more consistent with program expectations, and ongoing program inputs expected per theory were found to increase proportionally in classes observed. There was some continuing evidence of unruly student conduct (again, mostly at higher grade levels). However, compared to the first semester, the high rank and frequency of this irregularity had reduced reflecting only 16% of coded observations.

Participant Teacher Perceptions and Buy-in

Because the number is small, participant teacher survey results are not reported in a table. In the formative phase, only six of 9 participating teachers responded to program-related questions on the teacher survey (in the appendix, item-sections 36 and 38). Particularly, when asked if the program had any instructional value, all responding teachers opted to leave that item blank in the first semester.

At the end of the summative phase, there appeared to be greater acceptance of the *treatment* program by a majority of participating teachers compared to mid-year ratings. Notably, all the teachers responded to the survey. In all, 8 (89%) indicated that time for program management was "reasonable", given the supports they received; 7 (78%) indicated the content of the assignments was "effective"; 6 (67%) endorsed the "instructional value" of the program and found the worksheet format to be "effective"; and 7 (78%) indicated that time and other resource demands were "reasonable". Smaller numbers (1–5 of 9) of teacher participants chose "ineffective" responses to two questions or left them blank (11–56% respondents). These items dealt with time for providing individualized feedback, consistency of the supplemental program with regular curriculum (5, 56% positive responses in each), and other resource needs (4, 44% positive responses).

Table 5 is presented overleaf.

Table 5
Themes from Classroom Observations: Summary of Results from Formative and Summative Phases

Themes	Formative Phase Frequency (%)	Rank	Summative Phase Frequency (%)	Rank
1.0 Observed Inputs *Consistent* **with Program Theory**				
Materials— "diagnostic testing"; "tests" ; "testing materials"; "placement materials"; "daily assignments"; "supplies" "blue bins with materials organized"; "program shelves" "stacked materials";"game board"	61 (11%)	3	20 (9%)	5
Supports(developer/administration) — "staff" helping teachers assess; "staff" reading test directions; "staff" providing training in "school auditorium"; "aides" grading in program "resource room" ; "principal" stops in class; "principal inviting parents to orientation on phone"; "aides" helping teacher at table.	28 (4.8%)	7	34 (15%)	4
2.0 Helpful Inputs but *Inconsistent* **with Program Theory**				
"snacks" for children; "small class" sizes	5 (.09%)		None recorded	N/A
3.0 Observed Processes *Consistent* **with Program Theory**				
3.1 *Positive Environment*—teacher giving "positive reinforcement"; "children seated in small desk clusters as aide encourages one child to answer"; "diverse groups of children"; program "staff encouraging child"; "staff "helping"; teacher/aide "assisting" children.	84 (14%)	2	21 (9%)	5
3.2 *Teacher/Aides Following Protocol*— "giving directions"; "keeping time"; "grading assignments"; "monitoring make-up assessments"; "checking answers" "placing packets in bin"; "walking around": "looking at child's work".	56 (10%)	5	41 (18%)	2
3.3 *Students Engaged/following Protocol*— "self-checking answers" "placing packets back in bin"; "on-task"; "asking math questions"; "asking reading question"; "counting out loud when finishing sheet"; doing "assignments for the day".	32 (5%)	6	54 (24%)	1

Themes	Formative Phase Frequency (%)	Rank	Summative Phase Frequency (%)	Rank
4.0 Observed Processes *Inconsistent* with Program Theory				
4.1 *Teacher/Aide/Other Engaged in Student Behavior Management*—teacher/aide asking "students to be quiet"; "reprimanding students"; "sending students home"; "suspending students"; "loudly asking students to sit down".	61 (11%)	3	20 (9%)	5
4.2 *Student Misbehaviors during Supplemental Hour*—"playful"; "wandering around"; "noisy"; "distracted"; "complaining" "complaining loudly"; "not responding to directions"; "looking at others-not concentrating on work"; "chatting with others".	240 (41%)	1	35 (16%)	3
4.3 *Miscellaneous Activities outside Protocol and Program Plan*—"teacher moving children to a higher level worksheet before they reach mastery "; "Saturday sessions"; "parents picking up children before the supplemental hour has ended"; "teacher/aide letting children take worksheet home"; "children being moved out of XXX for disciplinary reasons".	15 (2%)	8	None recorded	N/A

$N_{codes\ (sem\ 1)}$ = 582; $N_{codes\ (sem\ 2)}$ = 225. Percentages are rounded.

Process-Outcome Relations: Formative Phase

Initially (Table 6–7), achievement outcomes were better for *treatment* children at the primary level rather than at the intermediate. Better outcomes were likewise found in reading than in mathematics, using the multi-level tests as mid-year outcome measures.

The combined primary level *treatment* group (Table 6) was 0.50 standard deviation (SD) units ahead of matched peers in mathematics performance, and 0.58 SD units ahead in reading performance. Although this difference was not statistically significant at the 5% error level, grade-level interactions were non-significant showing that the early influence of the supplemental program was similar in all primary grades. With grade level increases scores improved significantly in both groups.

In the combined intermediate grades (Table 7), *treatment* students were trailing behind their matched counterparts by -0.40 SD units in mathematics scores. This difference was significant at 10% error level (p=.08). In reading, Grade 5 students were 0.86 SD units ahead of matched peers while grade 4 students were -0.86 SD units below matched peers, generating an overall effect size of 0.035. The opposite results in Grades 4–5 yielded a significant interaction effect, showing that children in these two grade levels responded to the program differently (p<.01). The mixed achievement outcomes at the intermediate level could be stacked against observations gathered from the intermediate classrooms (Table 5) and attributed to the high levels of behavior problems documented.

Table 6
Results in Formative Phase: Reading and Mathematics Performance in Primary Students Receiving Supplemental Instruction

Outcome Variable	Mean	*SD*	R^2	Effect Size
Reading (Primary Level)				
Treatment	26.77	15.15	.342	0.58 NS
Comparison	22.22	7.83		
Mathematics (Primary Level)				
Treatment	83.68	27.76	.602	0.50 NS
Comparison	69.17	28.88		

ANOVA tables available on request.
NS not significant at 5% alpha level; F 1, 56 =2.42, p=.125
NS not significant at 5% alpha level; F 1, 52 =1.75, p=.192

Table 7
Results in Formative Phase: Reading and Mathematics Performance in Intermediate Students Receiving Supplemental Instruction

Outcome Variable	Mean	SD	R^2	Effect Size
Reading (Intermediate Level)				
Treatment	69.60	9.57	.237	0.04 NS
Comparison	69.29	8.78		
Mathematics (Intermediate Level)				
Treatment	54.32	23.11	.537	-0.40 *
Comparison	64.21	24.89		

ANOVA tables available on request.
NS not significant at p = .05; F 1, 45 =0.085, p=.772
* p < .10; F 1, 45 =3.305, p=.076

Teacher Perceptions and Treatment-Moderator Effects: Formative Phase

Table 4, referred to earlier, also showed the results on teacher-perceived levels of curriculum alignment, parent involvement and student independence in the classroom in the first and second phases of the investigation, based on means on teacher survey indices (see also the appendix for ratings on items 4–36). Findings were not very different over time or between *treatment* and *comparison* classroom teachers on composite survey indices. When means increased as they did on curriculum alignment with mathematics as the school year progressed, both *treatment* and *comparison* classroom teachers provided similar ratings on items, yielding comparable means. Comparison teachers reported marginally greater levels of Parent Involvement and Student Independence in their classrooms than *treatment* teachers.

Survey item-level ratings from the *summative phase* on skill-alignment (evidence of content validity of outcome measures in the appendix) were similar in both participating and non-participating classrooms, with greater levels of fit reported with reading curricula. In the reading area, close to 2/3 of 19 teachers in both programs indicated matches to a "great extent" between the supplemental program's reading skills and their curricula. In the math area, matches to a "great extent" were reported on recognizing numbers, reciting numbers, sequencing numbers, addition, and word problems (1/3 to 2/3 of teachers). The remaining math skill areas, such as subtraction, multiplication and division, generated very low proportions of positive ratings, even at the end of the year.

To check for moderating effects of differential levels of curriculum alignment, parent involvement or student independence in *treatment* and *comparison* classes, factorial ANOVAs showed that teacher-perceived curriculum alignment levels in reading in the primary sample had significantly different achievement effects in the formative phase (p=.05). Other results—a sampling of which is shown in the appendix—were non-significant for all other moderators in combined samples (primary and intermediate).The analyses were repeated in the summative study and the decision to use ANCOVAs was made after moderator effects were found to be non-significant.

Early Treatment Effects: Summative Phase.

Table 8–9 and Figure 3 show the results of the ANCOVAs. Overall, the *treatment* primary group was 0.45 standard deviation units ahead of *comparison* children in reading performance on skills/areas covered in the supplemental curriculum, unadjusted for mid-year performance (Table 7 and top two panels of Figure 3). Adjusted for mid-year scores, the *treatment* group was 3.4 raw units ahead. In combined primary grades, the *treatment* group was 0.58 standard deviation units ahead in mathematics performance. Adjusted for mid-year performance, the *treatment* students were still 5.09 raw score units higher than their matched counterparts. Although not statistically significant at the 5% error level, these effects may be classified as moderate in magnitude.

Table 8
Results in Summative Phase: Reading and Mathematics Performance in Primary Students Receiving Supplemental Instruction

Outcome Variable	Source of Variance (ANCOVA)	Mean Square	df	*F*	*p*
Reading at Primary Level	Regression (covariate)	6136.43	1	96.00	.000
	Treatment	172.29	1	2.70	.106
	Error	63.90	58		
Math at Primary Level	Regression (covariate)	1941.38	1	100.8	.000
	Treatment	116.58	1	0.6	.441
	Error	194.04	57		
Descriptive Statistics on Groups	Mean	*SD*	Adjusted Mean	R^2	Effect Size
Reading					
Treatment	31.1	13.1	30.7	.65	.45
Comparison	25.6	12.1	27.3		
Math					
Treatment	108.06	20.11	103.37	.70	.58
Comparison	93.20	26.20	98.28		

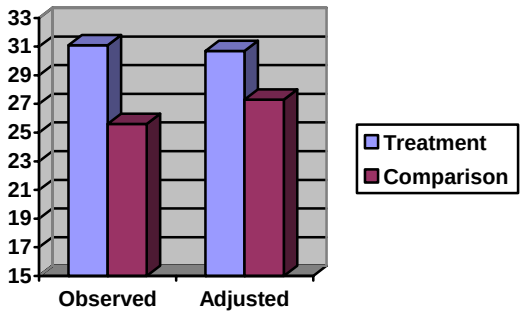
Covariate for both reading and math was the mid-year reading score; F for covariate* treatment interaction=1.93, p=.17 (reading);.F=0.12, p=.726 (math).

Table 9
Results in Summative Phase: Reading and Mathematics Performance in Intermediate Students Receiving Supplemental Instruction

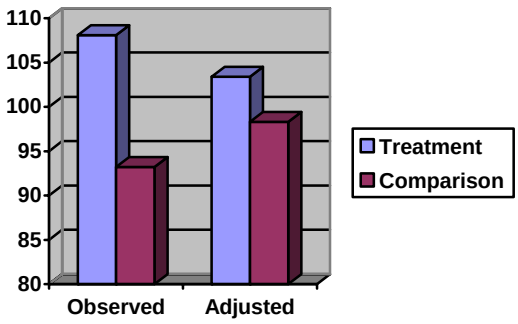
Outcome Variable	Source of Variance (ANCOVA)	Mean Square	df	*F*	*p*
Reading at Intermediate Level	Regression (covariate)	2501.27	1	15.83	.000
	Treatment	3.24	1	0.02	.887
	Error	157.99	44		
Math at Intermediatey Level	Regression (covariate)	9433.75	1	51.60	.000
	Treatment	1953.23	1	10.68	.002
	Error	182.82	50		
Descriptive Statistics on Groups	Mean	*SD*	Adjusted Mean	R^2	Effect Size
Reading					
Treatment	57.8	12.7	56.34	.27	.08
Comparison	56.5	17.2	57.15		
Math					
Treatment	84.1	21.3	84.17	.54	.65
Comparison	71.9	18.6	71.94		

Covariate for both reading and math was the mid-year reading score; F for covariate* treatment interaction=0.03, p=.863 (reading); F=.01, p=.968 (math).

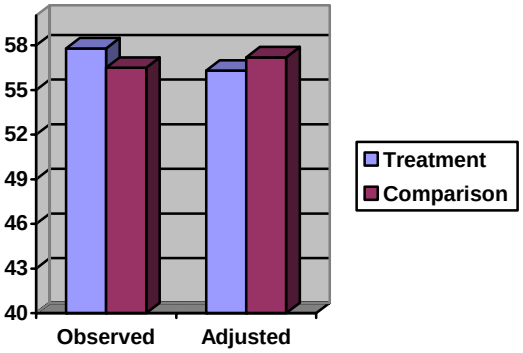
Treatment students were clearly ahead of their matched counterparts in the combined grade analysis at the intermediate level in mathematics (Table 8 and bottom panels of Figure 3), as evidenced in a positive effect size of 0.65 (p=.002). Adjusted for mid-year performance, the *treatment* students were still 12.23 raw units higher than their matched peers. In reading, however, there was a no discernable effect evident at the intermediate level (effect size of +0.08). Adjusted for mid-year scores, the *treatment* group was just 0.81 raw units below their matched peers.



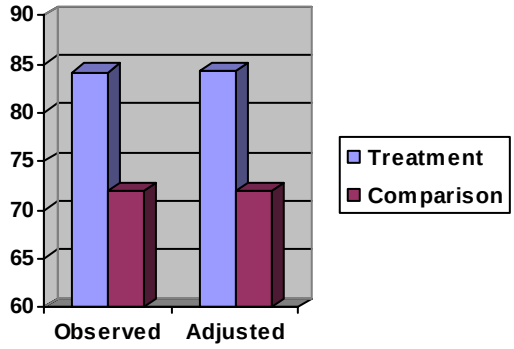
Reading Means in Primary Students



Math Means in Primary Students



Reading Means in Intermediate Students



Math Means in Intermediate Students

Figure 3. Summative ANCOVA Results: Effects of Supplemental Program on Achievement

Other Effects

Performance on district and state tests. On the Language Arts scale at the primary level, slightly higher proportions in the *treatment* group received teacher ratings of 5–6 (on a scale of 1–6) on Phonemic Awareness. In the other three areas, higher proportions of comparison students received ratings of 5–6. These differences were not statistically significant. On the CTB-4 math and reading test, the numbers of intermediate students with complete data changed from 2001 to 2002; thus these results could only be compared descriptively with 14 unmatched cases. They are not reported here due to instability of findings.

Test completion rates and time taken. Controlling for ranges of scores by quartile on the multilevel tests, a preliminary comparison of average time taken by students in *treatment* and *comparison* group suggested a pattern showing students who received supplemental services typically took 6–10 minutes less time to complete the tests. For example, the mean time taken in reading for students in the bottom quarter of the distribution was as follows at the primary level:

Table 10
Primary Time Required for Test Completion, Bottom Quartile, by Group

Group	Mean time taken (minutes)	*SD*
Treatment	25.4	3.4
Comparison	31.3	8.4

Controlling for grade level and given similar testing conditions, the mean number of items attempted by students was also higher in *treatment* classes in mathematics, and significantly different from comparison students ($F\ 1, 125 = 11.69$, $p<.001$). Typically, the *treatment* students attempted 2–6 more items at each grade in reading; in mathematics the average differences were approximately 8–20 more attempted items.

Self-efficacy measures. In the combined 4th and 5th grade samples, the *treatment* students had a mean *Math Self-efficacy* score of 23.0 (SD=4.0). The Comparison children had a mean of 24.3 (SD=3.8). This yielded an Effect Size of -.034, favoring the students *without* the Supplemental program. With the *Reading Self-efficacy* measure, the *treatment* students' mean was 18.6 (SD=3.3). The *comparison* children had a mean of 18.4 (SD=4.4), yielding an Effect Size of +.045, barely favoring the *treatment* students. Preliminary effects on self-efficacy were either absent or on the negative side.

To sum up, the early effects of the supplemental program were evident on skills tests aligned with the supplemental curriculum, but not on other measures. The developer and the school personnel were reminded that observed positive effects were "gross effects" and tentative; that is, results depicted the effects of the supplemental program as operationalized at the site and necessarily confounded with those of other reforms and supports concurrently aiming to raise student achievement. Confounders could not be teased out, as the program by its very definition was an add-on to the regular day programs in the same subject areas. However, the potential effects could still be broadly gauged in comparable groups to whom supplemental services were provided or withheld.

Discussion

The paper began with an aim to demonstrate and appraise a complete empirical application of the ETMM design for gathering research evidence on school-based programs and policy initiatives, in light of NCLB requirements calling for schools to implement programs supported by scientific evidence and the federal recognition of RCTs as the "gold standard" for scientific rigor. The focus was on a supplemental instruction program. The studies were done at one pilot site—an elementary school in Harlem.

At the outset, the reader should be reminded that the present ETMM application was limited by several field constraints and lack of resources, particularly, a time limit of one academic year. However, given these realities, what were the key advantages and disadvantages of the ETMM approach as compared to RCTs, had the latter been a design option under the same conditions? In the present application, the ETMM study was akin to small-scale, multi-method case study, focusing in-depth on implementation of a supplemental program at a particular site, and following the progress of the program as it matured and settled into a routine. It made inferences about possible early effects in treatment and comparison settings at one site only. A quasi-experiment was embedded in the design from the start, but formal linkages of program processes to outcomes were emphasized in the confirmatory phase of the research. Despite the time limit, there were *before* and *after* studies included in the investigation, driven by different purposes. Within the boundaries of one school, the study attempted a systemic approach to the design, making a formal effort to map and attend to the possible interactive/mediating effects of various *context*, *input*, *process* variables in the larger environment of a new program on *outcomes(CIPO)*. An analysis of a program's theory of action in terms of CIPO variables was thus the starting point of the design process.

As documented, several design challenges were faced once the studies were begun in the Harlem school. This is not uncommon in pilot efforts in real time school settings. Lessons were learned. Design changes were made— most design alterations were based on interactions with key stakeholders, formally gathered empirical evidence, and documented observations *in situ*.

Because of the use of comprehensive, mixed method approaches, there was better documentation of the various problems that arose in both treatment and comparison environments and the larger organization: sample attrition, emerging definition of the supplemental treatment in classrooms and the school, extent of treatment fidelity and stability as time passed, potential contaminants in the environment of both treatment and comparison students, such as student behavior problems. On all these, empirical data generated from the formative phase informed design decisions and changes. Because there were two separate phases of the research design, instrumentation issues could be tackled in the first phase with analyses of early impact held off until some evidence of validity and reliability was at hand on major variable measures. Stakeholders could look at the findings themselves and use the first phase results to alter program delivery; before-after comparisons could be made more meaningfully with an array of data from multiple sources. Teachers, leaders, parents gained more ownership of the new program by the second phase, improving delivery and fidelity.

Was it reasonable to incorporate a summative study within the pilot year of a new program? Ideally, the formative phase would last at least 1–2 years, with the summative phase starting soon after. Preferably, trained personnel would continue program implementation in the summative phase, either with cohorts students in the original treatment group continuing to receive services for studies of longitudinal effects, or with scaling up and expansion of the program to other, carefully selected sites to maximize generalizability and ecological validity of the confirmatory phase results. Scaled-up experiments using RCTs are best deferred until the second phase in ETMM studies; had

this been possible in the case presented here, it might have strengthened the quality of evidence (other things remaining constant). Feasible program models that emerged from the first phase could then be subjected to formal effectiveness testing in the second, using a tighter design that combined RCTs with other methods.

Questions may be raised about the ad-hoc instruments developed for the present ETMM application. A limitation was that early effects were evidenced only on specially-designed assessments specific to the supplemental curriculum and using the developer's item pool, rather than on independent, broader and standardized measures of achievement. Supplemental programs have narrower foci than regular curricula. When in pre-adoption stages, over-reliance on external standardized achievement measures may generate invalid findings due to issues of non-alignment/poor content validity. For optimizing local validity, instruments and data-gathering methods may thus need to be customized for small-scale testing and monitoring of novel programs, as shown here. At the same time, resources have to be dedicated to gathering sufficient evidence of validity and reliability for results to be defensible.

Several recommendations were made to developers and school personnel, with cautionary pointers on limitations. The developers were informed that increased alignment of a supplemental program with the regular-day curriculum's research base, content, and philosophy would likely improve outcomes as well as teacher and parent buy-in (as seen in teacher survey and student outcome data). The study also did not examine the quality of curriculum materials vis-à-vis the state's content standards and standards for best practices set by national subject area associations such as the National Council for Teachers in Mathematics and the National Council for Teachers of English. As necessary, developers were encouraged to examine the content of curricular products and their consistency with credible research, best practices, broader subject area domains tapped by national standardized achievement tests. Developers and school-based personnel were advised to plan program tryouts, replications, and related research with a longer term view, incorporating an understanding of the types of resources and conditions necessary for maximal success on particular outcome measures.

To compare the costs of the ETMM approach versus randomized field trials, the reader could weigh the breadth and quality of evidence generated from the present application versus the costs with RCT studies such as the 21st CLCC evaluation (described in the literature review). A main distinction is that the ETMM studies attend to program-development issues within particular environments while attempting to map a program's processes and effects over time. As shown in the present case, the smaller-scale ETMM design permitted more inclusiveness and participation of stakeholders and better relationship-building with researchers, making program improvements more likely. Despite the limitations, thus, the full-array findings were better understood through the documentation; stakeholders and researchers could appraise the results in a more informed manner—building trust amongst each other. In terms of disadvantages, the major design barrier of the ETMM application had to do with the high demands on resources and commitments of the developer, researchers, and sponsor to the project. Larger scale efforts could not be considered because of the intense human resource and material demands at a single site. These drawbacks must be weighed against the depth, meaningfulness and local utility of the body of information obtained.

How much better would the quality of evidence be if a traditional RCT had been implemented instead at the school described? Even if students had been randomly assigned to the supplemental services and control conditions at the start, the original RCT design would have been severely compromised because of factors such as teacher volunteers and high student mobility. With school-based innovations, thus, the answer may lie in carrying out a *small number* of in-depth, site-restricted, formative ETMM-type studies first. Once the first phase points to logistically feasible and promising program models ,a confirmatory phase could be initiated to scale up and test the models

with experiments . Such an approach may in fact be more cost-efficient in the long run than large scale randomized experiments (or quasi-experiments), without preparatory program-testing in natural settings. Compared to national implementations of RCTs, more limited and carefully-monitored ETMM-type field trials might better predict likely program impacts, and inform actions on subsequent program development and expansion.

In the end, the question as to how well ETMM designs compare with the federally-recommended gold standard must be left to the reader, other researchers, and users of research information. Further discussions should continue on alternate methods for improving scientific rigor of field studies and evaluations, particularly as successful instances of ETMM-type studies are documented in education and other fields.

References

Aguirre International (2000). *Save the children web of support initiative: Annual report 1999–2000*. San Mateo, CA: Author.

Anderson, L. M. & Weiner, L. (2004). *Early implementation of supplemental educational services under the No Child Left Behind Act: Year One Report*. Washington, D.C.: U.S. Department of Education, Office of the Under Secretary.

Bickman, L. (2000). Summing up program theory. *New Directions in Evaluation, 87*, 103–112.

Campbell, D. T. (1981). Introduction: Getting ready for the experimenting society. In L. Saxe & M. Fine. *Social experiments: Methods for design and evaluation* (pp. 13–18). Beverly Hills, CA: Sage Publications.

Chatterji, M. (2004). Evidence of "What Works": An argument for extended-term mixed method (ETMM) evaluation designs. *Educational Researcher, 33*(9), 1–13.

Chatterji, M. (2005). Reprint. Evidence of "What Works": An argument for extended-term mixed method (ETMM) evaluation designs. *Educational Researcher, 34*(6), 13–24.

Cohen, J. (2003). Supplemental education: Six essential components. *Principal, 82*(5), 34–37.

College Board. (1999). *Reaching the top: A report of the National Task Force on minority high achievement*. New York, NY: The College Entrance Examination Board.

Cook, T.D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175–199.

Cronbach, L. J., & Associates. (1980). *Toward reform in program evaluation*. San Francisco, CA: Jossey-Bass Publishers.

DeAngelis, K., & Rossi, R. (1997). *Schools serving family needs: Extended day programs in public and private schools*. Issue Brief. Washington, DC: American Institutes for Research. (ERIC Reproduction Document No. ED 406 022.)

Eisenhart, M. & Towne, L. (2003). Contestation and change in national policy on "scientifically-based" education research. *Educational Researcher, 32*(7), 31–38.

Harvard Family Research Project (2003). *Issues and opportunities in out-of-school time evaluation: Why, when and how to use evaluation—experts speak out*. Cambridge, MA: Harvard Family Research Project, Harvard Graduate School of Education.

Huang, D., Gribbons, B., Kim, K.S., Lee, C., Baker, E.L. (2000). *A decade of results: The impact of the LA's BEST after school enrichment initiative on subsequent student achievement and performance*. Los Angeles, CA: UCLA Center for the Study of Evaluation.

Mathematica Policy Research, Inc., & Decision Information Resources, Inc. (2003). *When schools stay open late: The national evaluation of the 21st Century Community Learning Centers Program*. Jessup, MD: U.S. Department of Education, ED Pubs.

No Child Left Behind Act of 2001 (NCLB), Public Law No. 107–110, 115 Statute 1425 (2002).

Ryan, G. W., & Bernard, H.R. (2000). Data management and analysis methods. In N.K. Denzin and Y.S. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 769–802). Thousand Oaks, CA: Sage.

U.S. Department of Education. (2002). *No Child Left Behind: The facts about 21st Century Learning*. Washington, DC: Author. Retrieved November 19, 2004, from http://www.ed.gov/pubs/21cent/firstyear.

University of California at Irvine, Department of Education (2001). *Evaluation of California's after-school learning and safe neighborhoods partnerships program: 1999–2000 Preliminary report*. CA, Irvine: Author.

Vallercamp, N., & Owens, D. (2003). Eight keys to a successful extended-day program. *Principal, 82*(5), 22–25.

Weischadle, D.E. (2002). Extended learning opportunities: Some lessons from the field. *Education, 123*(1), 73–81.

U.S. Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, D.C.: Institute for Education Sciences.

About the Author

Madhabi Chatterji
Teachers College, Columbia University

Young Ae Kwon
Kwon Learning Center, Charlotte, North Carolina

Clarice Sng
Teachers College, Columbia University

Email: mb1434@columbia.edu

Madhabi Chatterji, Ph.D., is Associate Professor of Measurement, Evaluation and Education at Teachers College, Columbia University. Her research interests are in evidence-gathering and evaluation of field interventions with systemic designs, designing classroom and institutional assessment systems, and in the design and validation of construct measures with classical and Rasch measurement methods.

Young Ae Kwon, Ed.D., is the owner and director of the Kwon Learning Center at Charlotte, North Carolina. Her research interests are in effective after-school programs for minority children and advanced statistical models.

Clarice Sng, Ed.D., is Associate Director, Office of Accreditation and Assessment, Teachers College, Columbia University. Her research interests are in instrument design and construct validation, mixed-method and longitudinal research.

Appendix: Supplemental Data

Supplemental Program Evaluation: Teacher Survey

Below is the text of the survey, with some ratings choices indicated in brackets.

> Purpose: This survey is intended for teachers whose classes are receiving the supplemental instruction program, as well as teachers of classes who are not. The purpose of this survey is to gather information on your classroom curriculum and environment, parent involvement levels, and if applicable, your current perceptions of the effectiveness and utility of the supplemental program.
> Time: The survey should take only 10–15 minutes to complete. Please respond to the questions as honestly and as thoughtfully as you can.
> Confidentiality: The results will be used in the study in aggregated form only. Although we are asking for individual teacher names or classroom identifiers for matching student names to correct classrooms, all the information will be coded anonymously and kept strictly confidential.
>
> THANKS FOR YOUR TIME
>
> Teacher Name:
> Classroom:
> Grade:
> Number of students:
> Room #:

Questions:

1. Does your classroom participate in the extended day supplemental program? [Yes/No]
2. Are students in your class repeating a grade and/or in a special education program? [Yes/No]
3. What innovative programs are in effect during the regular day in your classroom in reading and mathematics? (E.g., Success for all) List up to 3 key programs:

Curriculum Focus [reading]: To what extent are the following skills/areas addressed in the regular READING curriculum in your classroom? Use these responses(*Great Extent*, *Moderate Extent*, or *Little or Not at all*):

4. Reading comprehension in leveled passages
5. Listening comprehension in leveled passages
6. Identifying main ideas
7. Identifying details
8. Sequencing main ideas/details
9. Making connnections among ideas (e.g., cause and effect):

Curriculum Focus [Math]: To what extent are the following skills/areas addressed in the regular MATH curriculum in your classroom? Use these responses(*Great Extent*, *Moderate Extent*, or *Little or Not at all*):

10. Recognizing numbers
11. Reciting numbers
12. Sequencing numbers
13. Adding/subtracting 1–4 digit numbers in horizontal or vertical notation
14. Adding/subtracting 1–4 digit numbers with place value
15. Multiplication tables
16. Multiplication problems with 1–4 digits
17. Simple division
18. Long division
19. Word problems with above operations
20. Fractions
21. Adding fractions; subtracting fractions
22. Drawing lines/writing skills (Motor skills; hand-eye coordination)
23. List a maximum of 5 areas that you do emphasize that are not listed above:

Parent Involvement: Think of your class as a whole. To what extent are your students' parents/guardians involved in their student's education this year in the areas listed? Use these responses(*Great Extent*, *Moderate Extent*, or *Little or Not at all*):

24. Helping student with homework or academics
25. Responding to teacher requests/needs
26. Attending orientations/trainings
27. Attending school functions

Perceptions of Student Performance: Think of your class as a whole. Compared to the beginning of the year, to what extent are your students showing gains in these areas? Use these responses(*Great Extent*, *Moderate Extent*, or *Little or Not at all*):

Use these responses(A-C):

28. Mathematics
29. Reading
30. Writing (words, composing sentences, stories, themes)
31. Other subjects

Perceptions of Student Independence: Think of your class as a whole. Compared to the beginning of the year, to what extent are your students showing signs of self-directed and independent learning behaviors? Use these responses(*Great Extent*, *Moderate Extent*, or *Little or Not at all*):

32. Mathematics
33. Reading
34. Writing (words, composing sentences, stories, themes)
35. Other subjects

Supplemental Program Perceptions (RESPOND ONLY IF YOUR CLASS IS RECEIVING SUPPLEMENTAL SERVICES.) Rate your perceptions of the effectiveness of the program in these areas. [Scale: *Effective/Reasonable* or *Not Effective/Unreasonable*; comments also allowed.]

36. Quality of Worksheet Assignments
    - 36.1 Content
    - 36.2 Presentation/ format
    - 36.3 Consistency with regular curriculum
    - 36.4 Instructional value
37. Resource Needs
    - 37.1 Time for grading worksheets
    - 37.2 Time for program management

37.3 Time for providing individualized feedback
37.4 Other resource needs.
38. Program Support during pilot,

COMMENT ON WHAT WOULD NEED TO HAPPEN FOR YOU TO ADOPT THE PROGRAM.

Thanks again for your time!

Convergent Validity Evidence

Table A-1
Correlations of Multi-Level Reading and Math Composite Scores

Other variable	Reading Test Score	Math Test Score
Program Exposure (# of worksheets completed)	.62	.41
Number of Items attempted	.97	.70
Completion time	.92	.57

N=66; Split-half reliability ranges .67-.72 (primary); .78-.82 (intermediate)

Evidence of Content-validity of Multi-level Tests

Table A-2
Teacher Ratings of Curriculum Alignment of Supplemental Program (By Program)

Teachers' Survey Item	Supplemental Program Extent: Raw (%)			Comparison Program Extent: Raw (%)			Overall Program Extent: Raw (%)		
	Great	Moderate	Little /None	Great	Moderate	Little /None	Great	Moderate	Little /None
20. Fractions	5 (56)[a]	0	3 (33)	2 (20)	2 (20)	6 (60)	7 (37)[a]	2 (11)	9 (47)
21. Adding/ subtracting fractions	1 (11)[c]	1 (11)	4 (44)	0[b]	2 (20)	6 (60)	1 (5)[d]	3 (16)	10 (53)
22. Drawing lines/writing skills	4 (44)[a]	3 (33)	1 (11)	5 (50)[a]	3 (30)	1 (10)	9 (47)[b]	6 (32)	2 (11)

[a] One survey with no rating for this item; [b] Two surveys with no rating; [c] Three; [d] Five.

Table A-3
Teacher Ratings of Curriculum Alignment of Supplemental Program (By Program)

	Supplemental program Extent: Raw (%)			Comparison program Extent: Raw (%)			Both programs Extent: Raw (%)		
Teachers' survey item	Great	Moderate	Little /None	Great	Moderate	Little /None	Great	Moderate	Little /None
Curriculum alignment with Reading									
4. Reading comprehension	5 (56)	3 (33)	1 (11)	7 (70)	2 (20)	1 (10)	12 (63)	5 (26)	2 (11)
5. Listening comprehension	6 (67)	3 (33)	0	7 (70)	3 (30)	0 (0)	13 (68)	6 (32)	0 (0)
6. Identifying main ideas	5 (56)	4 (44)	0	7 (70)	2 (20)	1 (10)	12 (63)	6 (32)	1 (5)
7. Identifying details	5 (56)	4 (44)	0	6 (60)	3 (30)	1 (10)	11 (58)	7 (37)	1 (5)
8. Sequencing main ideas	4 (44)	5 (56)	0 (0)	6 (60)	3 (30)	1 (10)	10 (53)	8 (42)	1 (5)
9. Connections among ideas	5 (56)	3 (33)	1(11)	6 (60)	4 (40)	0	11 (58)	7 (37)	1(5)
Curriculum alignment with Math									
10. Recognizing numbers	8 (89)	1 (11)	0	9 (90)	1 (10)	0	17 (89)	2 (11)	0
11. Reciting numbers	6 (67)	2 (22)	1 (11)	7 (70)	3 (30)	0	13 (69)	5 (26)	1 (5)
12. Sequencing numbers	7 (78)	1 (11)	1 (11)	5 (50)	3 (30)	2 (20)	12 (63)	4 (21)	3 (16)
13. Adding 1–4 digit numbers	4 (44)	3 (33)	2 (22)	4 (40)	3 (30)	3 (30)	8 (42)	6 (32)	5 (26)
14. Adding 1–4 digit with place value	2 (22)	5 (56)	2 (22)	4 (40)	3 (30)	3 (30)	6 (32)	8 (42)	5 (26)
15. Multiplication tables	3 (33)[a]	2 (22)	4 (45)	0[a]	2 (20)	7 (70)	3 (16)[a]	4 (21)	11 (58)
16. Multiplication with 1–4 digits	2 (22)[a]	3 (33)	3 (33)	1(10)	2 (20)	7 (70)	3 (16)[a]	5 (26)	10 (53)
17. Simple division	1 (11)[a]	4 (44)	3 (33)	2 (20)	3 (30)	5 (50)	3 (16)[a]	7 (37)	8 (42)
18. Long division	0[a]	2 (22)	6 (67)	1 (10)	1 (10)	8 (80)	1 (5)[a]	3 (16)	14 (74)
19. Word problem with division	1 (11)[a]	3 (33)	4 (44)	5 (50)	0	5 (50)	6 (32)[a]	3 (16)	9 (47)

[a]—One survey with no rating for this item.

Self-efficacy Assessment Specifications and Sample Items

Table A-4
Theoretical Indicators and Matching Items in Math[a]

Domain Indicator Individual reports or displays a:	Sample Items	Response Scale
1. "Can do" Spirit and Belief in being Successsful in Subject	1. Can you finish your math work by yourself? 2. Can you complete the math work your teacher gives you? 3. When you see a new math problem, do you like to solve it by yourself? 4. Do you try to do your math work yourself before you ask for help?	a) Yes, all or most of the time b) Yes, some times c) No, very rarely or never
2. (Positive)Attitude towards Subject (no anxiety or fears)	5. Is doing math (number work) fun for you? 6. Do you like playing number games? 7. Do you like to work hard on math problems? 8. Do you think learning math will help you later?	a) Yes b) Unsure c) No
3. (Positive) Self-concept related to Subject	9. Are you good at math? 10. Have you always done well in math? 11. Do you think you get good grades in math? 12. Do you think you are just as good at math as your classmates?	a) Yes b) Unsure c) No

[a] Parallel Items were written for Reading and Mathematics.

Multi-level Test Specifications (Excerpts)

Table A-5
Ordered Content Indicators and Matching Items in Math[a]

Content Domain Indicator	Sample Items	Domain Weight	Level
1 Counting/sequencing Ordered indicators: Counting up to 10, counting to 20, counting to 100, recognizing and ordering object sets up to 5, up to 10, up to 20–30, sequencing numbers up to 10, up to 100, up to 200.	Item shows 5 objects on a page (e.g., pictures of 5 airplanes). Prompt says: *Count the pictures while pointing to each one*. Box provided for student to fill in number.	15%	Least difficult -- Primary
2 Addition and Subtraction Ordered indicators: Adding single digit numbers, adding two digit numbers, adding with numbers up to 100 with place value, adding of 2- and 3-digit numbers with place value (same indicators for subtraction)	185 + 325 ---------- 7-2 =	25%	Least difficult – Primary More difficult - Intermediate
3 Multiplication and Division Repeated addition, multiplication up to 3, multiplication up to 12, digits* 1-, 2-, 3- digits, 3digits*3digits (similar range of indicators for division, dividing with and without remainder)	___ x 8=48 185 x 5 ----------	25%	Least difficult – Primary Least to More difficult - Intermediate
4. Fractions Simple reduction, rewriting improper fractions, adding fractions with same or different denominators, Subtracting fractions	49/ 5= 2/5 + 3/5=	25%	Least to More difficult - Intermediate
5. Word problems Problems using simple operations; problems using more difficult operations	Tom had one cookie. Then Sue gave him 5 more. How many cookies did Tom have altogether?	10%	Least difficult – Primary More difficult - Intermediate

[a] Similar test design for reading domain.

Testing of Treatment x Moderators Effects on Outcomes: A Sampling of Results from the Formative Phase

Table A-6
Interaction Effects of Curriculum Alignment and Treatment on Mathematics

	Low Alignment		High Alignment		
Group	Mean	*SD*	Mean	*SD*	*N*
Treatment	71.94	26.48	68.74	31.14	53
Comparison	67.66	28.16	66.16	26.25	54
Source of Variance	Type III Sums of Squares	*df*	Mean Square	*F*	*p*
Treatment	297.751	1	297.751	0.367	0.546
Curr. Alignment	139.08	1	139.081	0.171	0.680
Treatment x Curr. Align.	18.35	1	18.359	0.023	0.881
Error	83615.54	104			

Table A-7
Interaction Effects of Parent Involvement and Treatment on Mathematics

	Low Involvement		High Involvement		
Group	Mean	*SD*	Mean	*SD*	*N*
Treatment	80.10	17.77	63.09	33.61	53
Comparison	76.44	25.93	58.79	25.66	54
Source of Variance	Type III Sums of Squares	*df*	Mean Square	*F*	*p*
Treatment	412.75	1	412.75	0.560	0.456
Parent Involvement	7828.47	1	7828.47	10.62	0.002
Treatment x Parent involvement	2.71	1	2.71	0.004	0.952
Error	75921.44	104			

See Table 3 and the survey described at the beginning of the appendix for survey indices and descriptive statistics. Median splits on survey indices were used to create sub-groups for both Tables A-6 and A-7. Levene's test for equality of variances was non-significant in all cases.

Table A-8
Interaction Effects of Student Independence and Treatment on Mathematics

	Low Independence		High Independence		
Group	Mean	*SD*	Mean	*SD*	*N*
Treatment	65.79	36.44	73.17	22.17	53
Comparison	55.42	26.82	76.20	23.82	54
Source of Variance	Type III Sums of Squares	*df*	Mean Square	*F*	*p*
Treatment	357.16	1	357.16	0.476	0.492
Student Independence	5247.88	1	5247.88	6.99	0.009
Treatment x Student independence	1188.42	1	1188.42	1.58	0.211
Error	77292.73	104			

See Table 3 and the survey described at the beginning of the appendix for survey indices and descriptive statistics. Median splits on survey indices were used to create sub-groups. Levene's test for equality of variances was non-significant in all cases.

EDUCATION POLICY ANALYSIS ARCHIVES **http://epaa.asu.edu**

Archivos Analíticos de Políticas Educativas