Evaluating the impact of an urban comprehensive school reform: An illustration of the need for mixed methods

Toni A. Sondergeld a,*, Kristin L.K. Koskey b,1

a Bowling Green State University, School of Educational Foundations, Leadership, & Policy, Education 517, Bowling Green, OH 43403, United States
b The University of Akron, College of Education, Educational Foundations & Leadership, Zook Hall 301N, Akron, OH 44325, United States

ABSTRACT

An abundance of comprehensive school reform (CSR) literature exists illustrating CSRs are effective in improving student outcomes. However, much of this research reports on top-down reforms, focuses on academic outcomes, and uses quantitative methods alone. Many educational researchers have argued for the use of mixed methods for providing a holistic understanding of a CSR’s impact on student outcomes. While mixed methods evaluations are frequently implemented, additional examples exemplifying “how” and “why” they should be used are needed. This study reports on the impact of an eight-year bottom-up CSR initiative on academic and non-academic student outcomes in an urban junior high school, and illustrates how the use of mixed methods was essential in providing a holistic evaluation of the program’s impact.

© 2011 Elsevier Ltd. All rights reserved.

Philosophy of comprehensive school reform

As part of the landmark No Child Left Behind Act of 2001 (NCLB), states, school districts, and schools are faced with increased accountability as a means of improving performance of all students in U.S. schools (U.S. Department of Education, 2004). Increased accountability refers to schools and districts meeting Adequate Yearly Progress (AYP) goals on standardized state tests in reading and math for the overall student population as well as historically underserved and underperforming subgroups. With the intent of closing the achievement gap between advantaged and disadvantaged students in reading and math, NCLB inadvertently “put predominately minority schools and racially integrated schools at a disadvantage in meeting their performance goals” (Sunderman, Kim, & Orfield, 2005, p. 26). This disadvantage is due to the fact that schools composed of primarily minority students (or a heterogeneous population) typically have AYP subgroup achievement targets (e.g., racial, economically disadvantaged, and limited English proficient) to reach in addition to the overall school AYP goal. As a result of needing to reach multiple goals, these schools are faced with the possibility of being struck with negative and potentially insurmountable sanctions for not meeting all AYP goals for the overall school as well as their minority subgroups. Thus, educators and administrators in economically disadvantaged and high minority schools are left feeling frustrated with the lack of direction provided by NCLB to meet the predetermined timeline for success (Hillner & Vance, 2006), and search for ways to help their schools succeed.

Comprehensive school reform (CSR) has become an increasingly popular method for such schools to attempt to meet state and nationally required standards and begin closing the achievement gap while providing students with high level instruction in classrooms. CSR efforts take a holistic approach where the philosophy of education of the school (or district) is brought into question and subject to change. The focus of CSR efforts is not only on high-stakes test outcomes, but also on non-academic outcomes such as attendance rate, behavior infractions, and student affect. Top-down or bottom-up forces, or a combination thereof, may be used to structure CSR efforts. Top-down reforms are more prescriptive in form and created by an external source such as school administrators, educational researchers, or private educational companies selecting or developing a CSR program to be implemented in teachers’ classrooms without teacher input. Bottom-up reforms are generated by schools alone or through educational partnerships where teachers have equal power in designing curriculum and reform efforts specifically based on their school’s needs. Combining these forces could mean a school uses an externally developed program selected by the school’s administration (top-down) and teachers modify it based on their own needs (bottom-up).

Previous CSR research has focused primarily on externally developed, or top-down, CSR efforts. Top-down reform efforts are strongly supported because they are easier to scale-up in order to reach larger numbers of the target audience (Datnow & Stringfield,
improvement of schools is possible when the reform effort is well thought-out, when teachers are active agents in the change process, when there are sufficient resources and time to support reform, when capable leadership is present, and when school cultures change along with school structures. These tenets about school change...have become common knowledge. (p. 184)

Additionally, it is clear that many different CSR programs have shown their effectiveness in improving student reading and/or math achievement across grade levels (e.g., Akay, 2006; Berends, Bodilly, & Kirby, 2002; Borman, Hewes, Overman, & Brown, 2002; Borman et al., 2006; Herlihy & Kemple, 2004; Munoz, Ross, & McDonald, 2007; Slavin & Madden, 2001; Zhang, Fashola, Skolnik, & Boyle, 2006). However, it is less evident through the research as to whether CSR programs have a clear impact on non-academic and affective domains simply because there is a lack of research conducted in these areas (see as exceptions Herlihy & Kemple, 2004; Munoz, Ross, & McDonald, 2007), even though they are of critical importance as well (Popham, 2004).

Despite the vastount of research conducted evaluating the impact of CSRs, much of this research has received scrutiny for its lack of rigor in design and inability to infer reliable program effects on achievement (Borman et al., 2002; Slavin, 2002). With the onset of NCLB, the bar was raised for school selection of educational reform programs as it became mandatory for schools “who depend on federal funding to select and implement programs that are based on scientific research” (Beghetto, 2003, p. 1). While the federal government’s view of scientifically based research (SBR) focused on utilizing experimental research designs alone, others such as the National Research Council argued that alternative research designs (e.g., qualitative, non-experimental quantitative, mixed methods) should also be included in the definition of SBR (Eisenhart & Towner, 2003). As a result of the federal call for SBR in education and lack of consensus as to what SBR means, the Comprehensive School Reform Quality (CSRQ) Center established standards for evaluating CSR programs including; design, assessment, timing, sampling, program implementation, and data analysis (Crowley & Hauser, 2007). To meet the CSRQ criteria, research designs should strictly be randomized controlled trials or quasi-experimental, and data analysis should employ appropriate statistical techniques thus enforcing the post-positivist nature of SBR as established by the federal government.

The need for mixed methods

Although the federal government (through NCLB) and the CSRQ suggest experimental methods are necessary to be considered SBR and infer causal relationships, others ardently fight for the value of additional methods. Rigorous quantitative research methods are indeed of value, and most everyone would agree with the importance of showing government funding monies are soundly spent (Feuer, Towner, & Shavelson, 2002). However, as previously argued, indicating these selective quantitative designs alone are the only methods of scientific inquiry capable of improving education has been widely rejected “because humans in schools are embedded in complex and changing networks of social interactions” (Berliner, 2002, p. 18). It is well recognized that the need for qualitative or mixed-methods inquiry is also of vital importance to better understand the underlying changes and complexities of individual situations (Maxwell, 2004; Pellegrino & Goldman, 2002). Instead of focusing on only experimental designs, critics often favor the “choice of method...linked to the question being studied” (Feuer, Towner, & Shavelson, 2002, p. 6). Thus, support has increased for utilizing a

Comprehensive school reform as a federally funded program

CSR was a federally funded program, thus requiring annual evaluations of program impact where implemented (Tushnet & Harris, 2006). Expanding on the concept of the Title I school-wide program first introduced in 1988, the Comprehensive School Reform Demonstration (CSR&D) was authorized in 1998 as Part F of Title I in the Elementary and Secondary Education Act and then signed into law as part of NCLB in 2002 (U.S. Department of Education, 2004). In 1998, Congress reserved 150 million dollars a year for the federal CSR&D program with the amount increasing dramatically shortly after to 310 million dollars a year (Slavin, 2002). Schools or districts awarded federal grants under the CSR&D program received a minimum of $50,000 a year for implementing a three-year program. Such monetary incentives caused for the rapid expansion of CSR efforts. By 2006, there were more than 700 different CSR&D programs being implemented. And since 1998, over 6800 schools nationwide have received funding for CSR from the federal government (Kidron & Darwin, 2007). Unfortunately, since 2007 the federal CSR&D program has not been supported politically or fiscally by the federal government (Borman, 2009). Despite this lack of support, CSR&D programs continue to thrive in our U.S. schools with funding being obtained through other avenues such as Title I money, Gear Up grants, and Race to the Top grants, which still require program effectiveness evaluations.

Evaluating the impact of CSR efforts

With government funding tied to many CSRs, a vast amount of CSR studies have been performed to establish how to most effectively implement a CSR program in order to raise student academic and non-academic outcomes. According to Datnow and Stringfield (2000), a substantive amount of research has been done on CSR programs to insure that
pragmatic perspective in program evaluation research where researchers are not wedded to any specific research paradigm, and alternatively implement both quantitative and qualitative methods when appropriate to yield a more holistic picture of “what” is occurring and “why” (Teddie & Tashakkori, 2009). Using such methods to provide a more complete picture is particularly important in evaluating CSRs given that the goal of CSR is to take a holistic approach to reform.

Applications of mixed methods in evaluation research to make educational decisions are nothing new (see, for a review, Greene, Caracelli, & Graham, 1989). Further, there are multiple existing frameworks to guide researchers in deciding the role of quantitative and qualitative data in evaluation research and how to integrate the two data types to inform a single phenomenon. For example, Greene and her colleagues (Greene, Caracelli, & Graham, 1989; Caracelli & Greene, 1993) have provided a framework to use in determining the purpose of using mixed methods in evaluation research. More specifically, Greene et al. (1989) outlined five main purposes for using mixed methods in evaluation research including: (1) triangulation (to examine the quantitative and qualitative data converge), (2) complementary (to use the quantitative and qualitative data to inform different aspects of a single phenomenon), (3) expansion (to use the quantitative data to expand the breadth and qualitative data to expand the depth of inquiry on a single phenomenon), (4) development (to use the quantitative or qualitative results to inform one another in making sampling decisions or survey development, for example), and (5) initiation (to use quantitative and qualitative methods to provide multiple perspectives on a single phenomenon to reveal contradictions). Additionally, Caracelli and Greene (1993) have provided a typology framework to use as a guide for conceptualizing ways to integrate quantitative and qualitative data during the data analysis stage. Example integration techniques are the use of data transformation (convert the quantitative or qualitative data into the other type of data for subsequent analyses) and data consolidation or merging (jointly review the quantitative and qualitative data to create a new variable for subsequent analyses). Others (e.g., Plano Clark, Garrett, & Leslie-Pelecky, 2010) have more recently expanded the discussion on advancing data integration techniques.

To assist in the ease of qualitative data analysis and data integration, advances in computer-assisted qualitative data analysis software (CAQDAS) programs have been made. Some of the more commonly used qualitative data analysis programs used to assist in organizing and analyzing large amounts of textual data include NVivo, ATLAS, and Ethnograph. QDA Miner is an additional program that provides for the analysis of both quantitative and qualitative data and thus facilitates the integration of the two forms of data. Reviewing the potential uses of such programs is beyond the scope of this article. We refer the reader to Peters and Wester (2006) for a more in-depth discussion on how CAQDAS can be used to assist in the analysis of qualitative data and to Leech and Onwuegbuzie (2007) for a brief discussion on the advantages and disadvantages of using such programs. Despite these and other advances in the mixed methods literature and the agreement on the importance of using mixed methods in evaluation research, there continues to be a call for additional examples of how quantitative and qualitative data can be integrated to inform a single phenomenon (Bryman, 2007; Cathain, Murphy, & Nicholl, 2007).

Purpose

Accordingly, the purpose of this study was twofold. First, the purpose was to evaluate the impact of a longitudinal bottom-up CSR effort in an urban junior high school on academic and non-academic student outcomes. Second, the purpose was to provide an illustrative example of how using quasi-experimental research alone in decision making for school reform evaluation is not always sufficient. We use this CSR evaluation to demonstrate that the implementation of a mixed methods design was essential for making meaningful inferences about the effectiveness of the bottom-up CSR program on academic and non-academic urban junior high school student outcomes. The results from this study illustrate how potential explanations of the quantitative results would not have surfaced without the qualitative phase. In the subsequent sections of this article, we first provide the context of the study and methods used. The quantitative and qualitative results are then reported. Next, the results are discussed within the context of the CSR literature. Finally, we discuss how the use of mixed methods was essential for providing a holistic understanding of the program's impact on student outcomes as well as the implications for evaluation research and, in turn, policy decisions.

Context of the study

One urban junior high school funded by a federal Gaining Early Awareness and Readiness for Undergraduate Programs (Gear Up) CSR for eight school years (2000–01 to 2007–08) was examined for changes in student outcomes (achievement, attendance rate, and behavior infractions). The main goal of this Gear Up CSR was to help provide support educationally disadvantaged students with educational choices they typically do not have access to make. This goal was to be accomplished through “accelerating student learning and establishing culturally responsive teaching practices designed to enhance social and academic preparation for college” (Kretovics, 2000, p. 4). Rather than using a traditional top-down model to reforming schools through a pre-established CSR program selected by administration or university faculty, a bottom-up approach was implemented. Reform in this junior high school started from the bottom-up where teachers were empowered to make decisions in the process of transforming their school based on their needs and the needs of their particular students (Kretovics, Farber, & Armaline, 1991). Specific plans to meet the overarching goals of the grant were co-created by the teachers involved in the school and professors at the university.

School reform at the gear up junior high school

Gear Up funding in the junior high school of interest began in the 2000–01 school year with teachers from the Gear Up junior high school meeting regularly after school for graduate level professional development (PD) courses with university professors. The primary objective of the first year’s PD offered to the teachers at the Gear Up school was to have these teachers research literature on urban education, best teaching practices, and school reform models. Teachers then used this information to generate their own CSR plan based on the needs of the students and community with whom they worked. The school-restructuring plan (the intervention or treatment in this study) derived by the Gear Up teachers included two main components: small school clusters and continuous professional development. Although a plan for school restructuring was established in the first year of the program (2000–01), the actual reform changes were not implemented until the second year of the program’s funding (2001–02). Further, the small school clusters and continuous professional development were adapted as needed based on the formative feedback provided during regular evaluation to meet the ongoing needs and improve the program to meet its goals. These components sometimes “shifted” in their focus based on the formative feedback during the evaluation process. However, these components were always a part of the program’s treatment and...
were modified to best meet the immediate needs of the students, teachers, and community.

**Small school clusters**

Small school clusters was one of the components of the school restructuring plan derived by the teachers of the junior high school examined in this study. The teachers chose to restructure into small school clusters based on the scientific research of Small Learning Communities (Cotton, 1996; Raywid, 1996). Such research “has convincingly demonstrated that small schools are superior to large ones on many measures and equal to them on the rest” (Cotton, 2001, p. 1). Some ways in which small schools have been shown to be superior over large schools include narrowing the achievement gap between ethnic minorities and white, middle class students and providing safer environments, and social and learning benefits.

In line with many of Sammon’s (2008) recommendations for creating effective small school clusters, the Gear Up teachers decided to restructure into small school clusters consisting of four core teachers (language arts, math, science, and social studies) working as a team with no more than 100 students per cluster. Teachers, rather than administrators or university professors, selected the core teacher clusters. Each teacher was then assigned, at random, 25 students from the cluster to mentor and assist with problems that might arise. In addition to individual daily planning time, core teachers had scheduled planning time to meet as a group and discuss curriculum, students, difficulties, and successes they encountered throughout the day. Specific curriculum was not provided with this CSR program; instead, a philosophy of high academic and non-academic expectations for all students was adopted. Teacher teams often planned integrated or thematic units to be taught across content areas to their students. Additionally, rewards and parties were planned for students who met academic and/or non-academic goals established by their team of teachers. Restructuring the school in small school clusters was intended to create small families of students and teachers thus leading to the improvement of student and teacher attitudes, attendance rate, behavior infractions, and academic achievement.

**Continuous professional development**

Another component of the school restructuring plan derived by the teachers of the junior high school examined in this study was continuous professional development. This component was created based on the understanding that structural changes alone “are inadequate to ensure improved student outcomes, a systematic, intensive, and long-term professional development program is needed for all staff involved in the education of students” (Kretovics, 2000, p. 29). Teachers at the Gear Up school were offered graduate level professional development courses on site. They could take these courses during common planning time and/or after school throughout the eight years of Gear Up in their school. Courses were not mandatory; therefore, teacher involvement in the professional development varied from one course to another based on teacher needs and/or interest.

Professional development began with teachers researching teaching and learning strategies, curriculum, instruction, urban education, and school reform models in the first year to develop their CSR model. Starting in year two, professional development courses for the Gear Up teachers focused on the needs identified by teachers rather than the assumptions of the university professors. Gear Up teachers collaboratively developed topics to be researched and studied in the graduate courses as well as the amount and duration of professional development. This bottom-up approach placed the professional development program’s “focus primarily on the teachers’ needs relative to developing curriculum units and instructional strategies tied to the cultural and background experiences of [their] students” (Kretovics, 2000, p. 29).

**Method**

**Research design**

A quasi-experimental case study using a sequential explanatory mixed methods design was employed in this research to investigate the efficacy of the Gear Up CSR. The quasi-experimental nature of this study is evident in that it had a treatment (Gear Up CSR program), outcome measures (achievement, attendance rate, and behavior infractions), and experimental units (students), but did not use random assignment of the experimental units (students) to infer if change was due to the treatment (Cook & Campbell, 1979). Further, the design of this study can be classified as a cohort design where successive junior high cohorts who received treatment (2001/02–2007/08) or no treatment (2000/01) at different times were compared (Crowley & Hauser, 2007) as matched samples. This research is also considered an intrinsic case study as our intent was to better understand this specific CSR program and the individuals (unit of analysis) involved (Stake, 1995).

While case studies are often referred to as a form of ethnographic research (Creswell, 2005), they may use qualitative, quantitative, or mixed methods (Yin, 1989). As illustrated in Fig. 1, a sequential explanatory mixed methods design was implemented in that the qualitative phase followed the quantitative phase (with priority given to the quantitative data) in order to further explain and better understand the overall quantitative findings observed (Creswell, 2005; Teddlie & Tashakkori, 2009). As the name suggests, explanatory mixed methods designs “are particularly useful for explaining relationships and/or study findings, especially when they are unexpected” (Hanson, Creswell, Clark, Petska, & Creswell, 2005, p. 229). For this reason, qualitative focus group interviews with purposefully selected school level administrators

![Fig. 1. Sequential explanatory design (illustration adapted from p. 154). Uppercase letters denote higher priority.](image-url)
and teachers were conducted after the quantitative analyses had been completed to inform the explanation of any statistically or practically significant quantitative findings or possible anomalies. Thus, specific research questions were not formulated for the qualitative component until after the quantitative phase was completed, characteristic of sequential explanatory mixed methods designs (Teddlie & Tashakkori, 2009).

Data collection and instrumentation

Quantitative phase

All quantitative data were received from the Ohio Department of Education or the school district’s data center. Attendance rate and total number of behavior incidents (in-school suspensions, out-of-school suspensions, and expulsions) per student data were collected and reported as end of school year results for non-academic outcome measures. Math and reading Ohio Achievement Test (OAT) data were obtained to measure academic outcomes.

Ohio Achievement Tests (OATs) are given to middle grades students in Ohio each spring over the content areas designated for their grade. Prior to the 2004–05 school year, valid assessments of student achievement did not exist in the state of Ohio as the old proficiency tests did not align well with what was actually taught in the classrooms. OATs were created with the purpose of rectifying this assessment problem by testing the state content standards that are taught in Ohio publicly funded schools. Content validity is therefore high as the assessments were created based on blueprints of the state content standards, which Ohio public school teachers are required to teach in their classrooms. While standardized tests may not document student learning well beyond the content of the tests, the OATs are well aligned with teacher classroom instruction. Additionally, standardized tests are a reliable indication of student learning (Kuncel & Hezlett, 2007), especially when compared to classroom grades that are often biased by teacher subjectivity (Malouff, 2008; Moore, 2002).

All OATs measure five levels of achievement including (from highest to lowest): Advanced, Accelerated, Proficient, Basic, and Limited. Regardless of the grade level or content, a scaled score of 400 is the lowest OAT Proficient score needed for passing. Individual tests, however, vary in their range, mean, standard deviation, standard error of measure, and reliability. Table 1 provides yearly detailed information on the test statistics resulting from 8th grade math and reading OATs. Table 2 illustrates the achievement level scaled cut score points for each test as they slightly vary between tests. Information for Tables 1 and 2 was obtained from the Ohio Department of Education website (www.ode.state.oh.us/) (ODE, 2008).

Population demographics for the Gear Up school are depicted in Table 3 for key factors over the eight years this study covered. A sample of 200 8th grade students was purposefully selected so they were proportionally representative of the overall average school demographics. Table 4 details the frequencies and percentages of students selected for each year.

Qualitative phase

A semi-structured interview protocol was purposefully created based on the findings from the quantitative phase. Specific interview questions about curriculum, programming, behavior, achievement, attendance, and atmosphere in the school were asked during semi-structured focus group interviews. The purpose of these focus group interviews was to help explain statistically and practically significant quantitative findings, and any unexpected findings. Figures illustrating the quantitative findings were provided to all participants at the time of the focus group interviews so each participant could easily view the quantitative results in question. The

### Table 1
8th grade yearly Ohio achievement test statistics by subject.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Max raw score</th>
<th>Max scaled score</th>
<th>Min scaled score</th>
<th>Raw score</th>
<th>Raw score SD (SEM)</th>
<th>Scaled score</th>
<th>Scaled score SD (SEM)</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring 2005</td>
<td>Math</td>
<td>142,783</td>
<td>46</td>
<td>551</td>
<td>282</td>
<td>19.35</td>
<td>9.10</td>
<td>29.45</td>
<td>30.00</td>
</tr>
<tr>
<td></td>
<td>Reading</td>
<td>143,044</td>
<td>48</td>
<td>539</td>
<td>258</td>
<td>29.54</td>
<td>9.48</td>
<td>40.47</td>
<td>30.01</td>
</tr>
<tr>
<td>Spring 2006</td>
<td>Math</td>
<td>141,290</td>
<td>46</td>
<td>534</td>
<td>290</td>
<td>21.30</td>
<td>9.35</td>
<td>31.51</td>
<td>26.75</td>
</tr>
<tr>
<td></td>
<td>Reading</td>
<td>141,369</td>
<td>48</td>
<td>548</td>
<td>267</td>
<td>26.46</td>
<td>9.58</td>
<td>42.09</td>
<td>31.18</td>
</tr>
<tr>
<td>Spring 2007</td>
<td>Math</td>
<td>139,826</td>
<td>46</td>
<td>550</td>
<td>278</td>
<td>21.39</td>
<td>8.43</td>
<td>31.67</td>
<td>25.63</td>
</tr>
<tr>
<td></td>
<td>Reading</td>
<td>139,748</td>
<td>48</td>
<td>538</td>
<td>260</td>
<td>29.54</td>
<td>9.07</td>
<td>42.66</td>
<td>28.33</td>
</tr>
<tr>
<td>Spring 2008</td>
<td>Math</td>
<td>135,630</td>
<td>46</td>
<td>546</td>
<td>291</td>
<td>22.38</td>
<td>9.40</td>
<td>31.85</td>
<td>28.94</td>
</tr>
<tr>
<td></td>
<td>Reading</td>
<td>135,719</td>
<td>48</td>
<td>543</td>
<td>259</td>
<td>28.01</td>
<td>9.39</td>
<td>42.19</td>
<td>29.33</td>
</tr>
</tbody>
</table>

### Table 2
8th grade achievement level scaled cut score points.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Limited</th>
<th>Basic</th>
<th>Proficient</th>
<th>Accelerated</th>
<th>Advanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>&lt;378</td>
<td>378</td>
<td>400</td>
<td>428</td>
<td>451</td>
</tr>
<tr>
<td>Math</td>
<td>&lt;379</td>
<td>379</td>
<td>400</td>
<td>432</td>
<td>459</td>
</tr>
</tbody>
</table>

### Table 3
Demographic data for the gear up school by year.

<table>
<thead>
<tr>
<th>Year</th>
<th>Total enrollment</th>
<th>Minority students</th>
<th>Economically disadvantaged</th>
<th>Special education</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000–2001</td>
<td>809</td>
<td>29.0%</td>
<td>49.5%</td>
<td>19.9%</td>
</tr>
<tr>
<td>2001–2002</td>
<td>859</td>
<td>31.7%</td>
<td>56.6%</td>
<td>18.5%</td>
</tr>
<tr>
<td>2002–2003</td>
<td>804</td>
<td>34.1%</td>
<td>52.7%</td>
<td>17.4%</td>
</tr>
<tr>
<td>2003–2004</td>
<td>743</td>
<td>32.5%</td>
<td>53.3%</td>
<td>18.7%</td>
</tr>
<tr>
<td>2004–2005</td>
<td>731</td>
<td>35.9%</td>
<td>63.2%</td>
<td>19.5%</td>
</tr>
<tr>
<td>2005–2006</td>
<td>695</td>
<td>35.5%</td>
<td>81.9%</td>
<td>18.7%</td>
</tr>
<tr>
<td>2006–2007</td>
<td>682</td>
<td>36.0%</td>
<td>80.7%</td>
<td>18.7%</td>
</tr>
<tr>
<td>2007–2008</td>
<td>715</td>
<td>38.9%</td>
<td>77.1%</td>
<td>21.3%</td>
</tr>
<tr>
<td>Avg. across years</td>
<td>755</td>
<td>34.2%</td>
<td>64.4%</td>
<td>19.1%</td>
</tr>
</tbody>
</table>
The interviewer described the statistical/practical/unexpected findings and asked the participants to explain from their perspective why the quantitative research findings may have occurred in order to refine the understanding of the results. This procedure is typical of explanatory mixed methods research where the interview questions are generated after the quantitative phase is complete to provide a more in-depth understanding of what is observed in the quantitative results. Appendix A provides an example of this process including the figures and interview questions that were used to ask teachers about the quantitative findings on behavior incidents over time in this study. Prior to utilizing the interview protocol two experts in the field of educational research and evaluation reviewed and helped to modify the protocol for clarity and completeness. In addition to the focus group interviews, Gear Up activity reporting documents collected over the length of the program's funding for federal reporting purposes were used to corroborate teacher/administrator responses whenever possible. Activity reporting documents identified the numerous Gear Up funded programs, number of student participants, and duration of the activities for each school year.

Three semi-structured focus group interviews with teachers and administrators lasting approximately 30 min each were conducted. Participants were purposefully selected based on their position in the school and the content they taught. Seventeen teachers and administrators were interviewed: 8 teachers taught Language Arts, 6 taught Math, 1 taught physical education (former Gear Up liaison), and 2 were school-level administrators. Gender and ethnicity were unevenly distributed with 14 females and 3 males, 14 White and 3 Black. Total teaching experience ranged from 4 to 33 years and years teaching in this particular school ranged from 1 to 33 years.

**Data analysis**

**Quantitative phase**

Four repeated measures (within-subjects) analysis of variance (ANOVA) tests were performed to test for significant differences in the dependent variables (achievement, attendance rate, behavior infractions) over time using Statistical Packages for the Social Sciences (SPSS). Within-subjects designs are advantageous to use as they require fewer participants than between-subjects designs, and are more statistically powerful because the error variance attributed to individual variation is removed (Field, 2005; Lamb, 2003). We used a within-subjects design rather than a between-subjects design due to the fact that no suitable comparison group with similar key demographics in the same school district existed. For repeated measures designs with more than two levels, SPSS automatically generates output for Mauchley's Test of Sphericity. If this test result were statistically significant ($p < 0.05$) in this study, either the Greenhouse-Geisser or Huynh-Feldt correction was employed. The Greenhouse-Geisser correction is known for being too conservative when sphericity is high, and the Huynh-Feldt correction is known for overestimating sphericity (Field, 2005). Based on the limitations of each sphericity correction, when estimates of sphericity were greater than 0.75, the Huynh-Feldt correction was used, and when estimates of sphericity were less than 0.75 the Greenhouse-Geisser correction was used (Field, 2005). If a statistically significant effect for time was found, further pairwise comparisons with a Sidak adjustment for multiple comparisons were used to assess where the specific statistically significant differences were between consecutive years.

**Qualitative phase**

Focus group interviews were transcribed immediately following the interviews. A thematic analysis (Miles & Huberman, 1994) of the transcriptions was performed. Sorting of the qualitative data for this study was done in two steps. First, data were placed into broad categories based on which quantitative research findings they helped to inform (e.g., Math OAT results, Attendance Rate results, etc.). Some qualitative data assisted with the understanding of more than one quantitative finding and were therefore placed in multiple categories. Next, finer themes within each of the initial broad categories were identified and used to further explain "why" the quantitative results may have occurred.

**Data integration**

Integration of the quantitative and qualitative data is key when conducting mixed methods research (O’ Cathain, Murphy, & Nicholl, 2007; Teddile & Tashakkori, 2009). In this study, we integrated the data in three ways. First, we used a form of what Caracelli and Greene (1993) refer to as extreme case analysis where extreme cases or anomalies in either the quantitative or qualitative data are identified and further pursued using the other method for further clarification or explanation. In this case, we identified anomalies in the quantitative data and conducted additional qualitative data collection and analyses seeking explanation for these anomalies. Second, we integrated the quantitative and qualitative data when reporting the results on the academic and non-academic outcomes. Third, the quantitative and qualitative data were used to make meta-inferences regarding the impact of the CSR on these outcomes.

**Results**

Since the mixed methods approach to this study was explanatory in nature, quantitative findings for each dependent variable of interest are shared first, followed by qualitative findings from the focus group interviews that helped to further clarify any statistically and/or practically significant quantitative findings and/or anomalies. Specific frequencies of teachers and administrators supporting the qualitative results cannot be identified for many of the findings due to the nature of conducting focus group interviews. However, all teacher and administrator beliefs shared during the interviews are represented.

**Achievement results**

**Quantitative math achievement results**

A 4-Within RM-ANOVA was used to test for differences in 8th grade math achievement scores in the Gear Up students from the 2004–05 school year to the 2007–08 school year. Mauchly’s Test of Sphericity was statistically significant ($p < 0.001$) and Epsilon was greater than 0.75, therefore a Greenhouse-Geisser correction was applied. Eighth grade math achievement scores significantly changed over time as the within-subjects effect for time was statistically significant, $F(2.65) = 209.53$ ($p < 0.001$). The effect size$^2$

---

$^2$ The following criteria from Cohen (1968) were used to interpret the strength of the effect sizes: small = $1$%, medium = $10$%, large = $25$%.
for time was large with 51.3% of the variability in 8th grade math OAT scores being attributed to the school year. Table 5 depicts the means and standard deviations for each year's 8th grade math OAT scores and Table 6 shows results from the RM-ANOVA.

Subsequent pairwise comparisons with a Sidak adjustment for multiple comparisons revealed that on average, 8th grade math OAT scores increased 22.87 points from 2004–05 to 2007–08 with statistically significant score increases each year. The statistically significant increase in average 8th grade math OATs raised the average score from below proficient in 2004–05 through 2006–07 (≤400) to above proficient (>400) in 2007–08. Table 7 shows the results from the paired comparisons and Fig. 2 displays the trend over time graphically.

Qualitative math achievement results

Teachers and administrators were told by the interviewer that 8th grade OAT math scores significantly increased over time. However, on average, the scores were between 15 and 20 points below the 400 proficient level in the 2004–05 school year and the 2005–06 school year. Dramatic increase to average 8th grade math OAT scores to approximately proficient occurred in the 2006–07 and above proficient in the 2007–08 school years. Both administrators ardently believed the significant increase in scores was largely due to teacher turnover in the math department. One administrator clearly indicated this as she stated “There has been a big turnover in math staff and there are really good teachers there now. But, in the 2004–05 and 2005–06 school years there were really bad math teachers in the school.”

Current Gear Up math teachers felt Title I services significantly helped their student scores increase from the 2006–07 school year to the 2007–08 school year. Eighth grade math teachers reported that “almost half of the students in the 8th grade received an extra math class a day with a Title I math teacher” in the 2007–08 school year. This meant that nearly half of their students had a total of two full math classes a day.

Math teachers also suggested that Gear Up professional development had assisted in improving their teaching methods. Although some of these teachers had gone to specific math professional development courses offered through Gear Up (e.g., Connected Mathematics training), most of these teachers believed the technology (e.g., Smart Boards, Making the Grade, Inspiration, United Streaming, etc.) and behavior management courses had the greatest impact on their teaching and interaction with their students. Additionally, they considered the curriculum mapping they performed with the Gear Up professors to be very beneficial and felt it led to “better and more cohesive instruction” that “helped keep teachers [instructionally] on the same page.”

Quantitative reading achievement results

A 4-Within RM-ANOVA was used to test for differences in 8th grade reading achievement scores in the Gear Up students from the 2004–05 school year to the 2007–08 school year. Mauchly’s Test of
Sphericity was statistically significant \((p < 0.001)\) and Epsilon was greater than 0.75, therefore a Greenhouse-Geisser correction was applied. Eighth grade reading achievement scores significantly changed over time as the within-subjects effect for time was statistically significant, \(F(2.60) = 36.13\ (p < 0.001)\). The effect size for time was large with 15.4% of the variability in 8th grade reading OAT scores being attributed to the school year. Table 5 depicts the means and standard deviations for each year’s 8th grade reading OAT scores and Table 6 shows results from the RM-ANOVA.

Subsequent pairwise comparisons with a Sidak adjustment for multiple comparisons revealed that on average, 8th grade reading OAT scores increased 9.13 points from 2004–05 to 2007–08. The scores remained approximately the same, nearly proficient (400), from 2004–05 to 2005–06. In 2006–07 and 2007–08 8th grade reading OAT scores significantly increased each year raising the average score above the minimum proficient level of 400. Table 7 shows the results from the paired comparisons and Fig. 3 displays the trend over time graphically.

**Qualitative reading achievement results**

Teachers and administrators were asked why they felt reading OAT scores were nearly proficient and then significantly increased to above proficient over the course of this study. Both Gear Up school administrators attributed the significant increase in 8th grade reading OAT scores over time to the “superior” or “just excellent” language arts teachers. Further, they also believed it was helpful that there had been “far less teacher turnover in the language arts department” compared to the math department. And when new teachers entered the language arts department, “veteran teachers have brought the new teachers into their clusters by welcoming them and getting them on board with their plans.” The 8th grade language arts teachers and the administration agreed that the language arts department as a whole had become a “tight knit” group and Gear Up helped them to improve their “sense of community.”

The 8th grade teachers reported their strategy over the years had been to have students “read, read, read every day.” They have “used benchmark assessments at the beginning of the year to see where students are with their reading ability and then go from there.” Without the classroom sets of novels and books Gear Up had purchased for them, these teachers felt this strategy would have been difficult if not impossible to implement since they “simply did not have the materials to do this before.”

Language arts teachers also indicated Gear Up professional development and funding helped improve their individual and collective quality in several ways. First, the Gear Up led curriculum mapping assisted greatly in that teachers “know where they are” going during the year and sometimes do cluster cross-curricular units together.” They felt this “global plan” was quite useful and therefore re-amped their curriculum maps over the summer of 2006 even though Gear Up professors were no longer working with them as the grant funding and support was gone at this point. Second, Gear Up funding allowed for the purchase of “lots of books for classrooms which they use to read a lot.” Gear Up funding also supported a program called “Reading Quest”—an afterschool book club that had very high participation (more than 50 students). This reading enrichment program rewarded students with field trips to places such as Cleveland, Ohio and Cedar Point (a local amusement park) “for showing up to the program a certain amount of times.” Reading Quest was eliminated when the Gear Up funding left because the school simply could not afford to financially support the program on their own.

**Attendance rate results**

**Quantitative attendance rate results**

An 8-Within RM-ANOVA was used to test for differences in attendance rate in the Gear Up students from the 2000–01 school year to the 2007–08 school year. Mauchly’s Test of Sphericity was statistically significant \((p < 0.001)\) and Epsilon was less than 0.75, therefore a Huynh-Feldt correction was applied. Attendance rate significantly changed over time as the within-subjects effect for time was statistically significant, \(F(3.57) = 482.83\ (p < 0.001)\). The effect size for time was large with 54.8% of the variability in attendance rate being attributed to the school year. Table 5 depicts the means and standard deviations for each year’s overall attendance rate and Table 8 shows results from the RM-ANOVA.

Subsequent pairwise comparisons with a Sidak adjustment for multiple comparisons revealed that on average, overall attendance rate significantly increased over time from 2000–01 to 2007–08 by 8.74 percentage points from averaging 85.84% (below the state AYP benchmark) to 94.58% (above the state AYP benchmark), respectively. Table 9 shows the results from the paired comparisons, and Fig. 4 displays the trend over time graphically.

**Qualitative attendance rate results**

When questioned as to why teachers and administrators believed students attendance rate had significantly increased over time from averaging approximately 86% pre-Gear Up to 95% in the last year of Gear Up all teachers and administrators appeared somewhat “surprised” to see an increase of this magnitude. Rather than attributing the rise in student attendance strictly to Gear Up, teachers and administrators felt it likely had more to do with change in state reporting policy and questionable ability of school attendance.
faculty. With regard to policy, disciplinary Alternative Placements had previously been counted as student absence and recently were not. Alternative Placements are offered to students who receive an out of school suspension. These students go to a different building to complete assignments as an alternative to missing their work and being counted absent. Consequently, some teachers agreed that “attendance is a big game” with the rules regularly changing.

Teachers and administrators also felt strongly that attendance rate was largely dependent on “who was keeping track of it and who was reporting it.” According to an administrator, attendance secretary “is the least desirable position in the building” and has in the past “not been reported correctly.” A teacher reported that up until the 2008–09 school year there were often “students who were actually in the school marked as absent and students not in school not reported absent on the absent sheet.”

The teachers did feel that attendance may have increased somewhat due to the fact that “Gear Up helped the school to become teacher-led and we now look at attendance numbers more.” Further, teachers felt that they “now talk to students and parents more about where students are when they’re not in school and why they’re out. We just show more interest in this aspect by calling home and sending notes when students are absent.” This change reportedly began when they formed teacher clusters for Gear Up, and the clusters started working together to form communities for the teachers and students.

Behavior infraction results

Quantitative behavior infraction results

An 8–Within RM-ANOVA was used to test for differences in behavior infractions in the Gear Up students from the 2000–01 school year to the 2007–08 school year. Mauchly’s Test of Sphericity was statistically significant (p < 0.001) and Epsilon was greater than 0.75, therefore a Greenhouse-Geisser correction was applied. Behavior infractions significantly changed over time, as the within-subjects effect for time was statistically significant, F(5.39) = 73.83 (p < 0.001). The effect size for time was large with 15.6% of the variability in behavior infractions being attributed to the school year. Table 5 depicts the means and standard deviations for each year’s overall average behavior infractions per student and Table 8 shows results from the RM-ANOVA.

Subsequent pairwise comparisons with a Sidak adjustment for multiple comparisons revealed that on average reported behavior infractions significantly decreased from an average of 2.09 per student in 2000–01 to an average of 1.42 per student in 2007–08. An anomaly in this trend was revealed when average behavior infractions per student significantly increased in 2005–06 to 2.27, which was greater than that which was reported in the baseline year (2.09). In 2006–07, the average behavior infractions per student significantly decreased, returning to statistically the same level as 2004–05 (approximately 1.60), which was before the 2005–06 behavior incident spike. Table 9 shows the results from the paired comparisons, and Fig. 5 displays the trend over time graphically.

Qualitative behavior incidents results

Two main questions needed clarification by teachers and administrators with regard to quantitative behavior incident results. First, they were asked for an explanation of why behavior incidents had improved by significantly decreasing over time. Second, it was imperative to gain insight from the teachers and administrators as to why there was a statistical anomaly in the 2005–06 school year with behavior incidents deviating from the trend and spiking above.

Table 8
Repeated measures analysis of variance for non-academic outcomes.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>$\eta^2$</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendance rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>20,944.40</td>
<td>3.58</td>
<td>5854.96</td>
<td>482.83***</td>
<td>0.55</td>
<td>2.12</td>
</tr>
<tr>
<td>Error</td>
<td>17,308.16</td>
<td>1427.31</td>
<td>12.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>38,252.56</td>
<td>1430.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of behavior infractions/student</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>234.42</td>
<td>5.39</td>
<td>43.53</td>
<td>73.83***</td>
<td>0.156</td>
<td>0.86</td>
</tr>
<tr>
<td>Error</td>
<td>1266.95</td>
<td>2148.66</td>
<td>0.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1501.37</td>
<td>2154.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*** p < 0.001.

Table 9
Pairwise comparisons between years for non-academic outcomes.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendance rate (I)</td>
<td>0.84*** (0.18)</td>
<td>−5.21*** (0.20)</td>
<td>−4.55*** (0.18)</td>
<td>−3.69*** (0.16)</td>
<td>−5.08*** (0.19)</td>
<td>−4.92*** (0.23)</td>
<td>−8.74*** (0.28)</td>
</tr>
<tr>
<td>2001–02</td>
<td>−</td>
<td>−</td>
<td>−0.92*** (0.20)</td>
<td>−2.86*** (0.14)</td>
<td>−4.24*** (0.15)</td>
<td>−4.08*** (0.20)</td>
<td>−7.90*** (0.27)</td>
</tr>
<tr>
<td>2002–03</td>
<td>−</td>
<td>−</td>
<td>0.75*** (0.12)</td>
<td>1.60*** (0.13)</td>
<td>0.21 (0.11)</td>
<td>0.38 (0.14)</td>
<td>−3.45*** (0.17)</td>
</tr>
<tr>
<td>2003–04</td>
<td>−</td>
<td>−</td>
<td>−0.85*** (0.13)</td>
<td>0.53*** (0.13)</td>
<td>−0.37 (0.17)</td>
<td>0.19*** (0.20)</td>
<td></td>
</tr>
<tr>
<td>2004–05</td>
<td>−</td>
<td>−</td>
<td>−1.39*** (0.12)</td>
<td>−1.22*** (0.15)</td>
<td>0.16*** (0.13)</td>
<td>0.55*** (0.05)</td>
<td></td>
</tr>
<tr>
<td>2005–06</td>
<td>−</td>
<td>−</td>
<td>−0.44*** (0.18)</td>
<td>−1.82*** (0.18)</td>
<td>3.82*** (0.18)</td>
<td>6.66*** (0.18)</td>
<td></td>
</tr>
<tr>
<td>2006–07</td>
<td>−</td>
<td>−</td>
<td>−0.31*** (0.05)</td>
<td>0.36*** (0.05)</td>
<td>0.32 (0.05)</td>
<td>0.49*** (0.05)</td>
<td></td>
</tr>
<tr>
<td>Number of behavior infractions/student</td>
<td>0.12 (0.04)</td>
<td>0.18 (0.05)</td>
<td>0.40*** (0.05)</td>
<td>0.49*** (0.05)</td>
<td>0.18*** (0.04)</td>
<td>0.50*** (0.05)</td>
<td>0.67*** (0.05)</td>
</tr>
<tr>
<td>2000–01</td>
<td>−</td>
<td>−</td>
<td>−0.06 (0.04)</td>
<td>0.28*** (0.04)</td>
<td>0.37*** (0.04)</td>
<td>−0.30*** (0.05)</td>
<td>0.38*** (0.05)</td>
</tr>
<tr>
<td>2001–02</td>
<td>−</td>
<td>−</td>
<td>0.22*** (0.05)</td>
<td>0.31*** (0.05)</td>
<td>0.36*** (0.05)</td>
<td>0.32 (0.05)</td>
<td>0.49*** (0.05)</td>
</tr>
<tr>
<td>2002–03</td>
<td>−</td>
<td>−</td>
<td>−0.08 (0.04)</td>
<td>0.057*** (0.05)</td>
<td>−0.59*** (0.05)</td>
<td>0.10 (0.05)</td>
<td>0.27 (0.04)</td>
</tr>
<tr>
<td>2003–04</td>
<td>−</td>
<td>−</td>
<td>−0.67*** (0.06)</td>
<td>−0.02 (0.04)</td>
<td>0.19*** (0.04)</td>
<td>0.19*** (0.04)</td>
<td></td>
</tr>
<tr>
<td>2004–05</td>
<td>−</td>
<td>−</td>
<td>−0.67*** (0.06)</td>
<td>0.67 (0.06)</td>
<td>0.85*** (0.06)</td>
<td>0.17 (0.04)</td>
<td></td>
</tr>
</tbody>
</table>

Sidak adjustment made for multiple comparisons.

*** p < .01.

** p < .001.

Behavior Plans reportedly enabled teachers to take more “steps” within their classrooms and clusters to prevent discipline referrals which had increased in the past led to suspensions. For example, an administrator reported that “one cluster even has 10 steps their students must go through before they get a referral. So the teachers feel more in control and they are taking care of more minor problems than they ever did before.” Teachers indicated that they were also “talk[ing] more about what’s going on each day, and fix[ing] behavior issues immediately between classes since [teachers in the same] clusters are now physically located near each other.” Although student behavior may not have been improving dramatically from year to year as far as the Gear Up teachers and administrators believed, the way in which negative behavior was handled drastically changed with teachers managing more of the discipline issues and less incidents being reported as referrals to the administration. Additionally, while Gear Up funding was in the school (2001–2008), many Gear Up funded incentives (e.g., field trips to professional baseball games, picnics, bowling parties, movies, etc.) for good behavior were implemented on a quarterly basis which teachers believed promoted better behavior among many of their students.

Reported behavior incidents in the Gear Up school decreased every school year, suggesting behavior was improving in the school with the exception of the 2005–06 school year. In the 2005–06 school year, negative behavior spiked, rising above the greatest reported number of negative behavior incidents in the 2000–01 school year prior to Gear Up implementation. Teacher and administration perspectives on the anomaly of this significant behavior increase were considerably different. In the 2005–06 school year, teachers were informed in late February 2006 that their school building would be closing, and no one knew where the students and teachers would “be going the next year and if they would all be staying together or not.” The teachers who had been in the school for longer periods of time remembered the middle and end of this school year being very “unsettled” with great “chaos and animosity felt by teachers and parents.” They believed that this “unpleasant atmosphere” impacted the students as well, where the students “did not seem to care anymore once they knew the building was closing.” Administrators, however, did “not recall the 2005–06 school year being a particularly bad year.” They did not recognize “any particular unsettlement within the school that would make the students behave any differently than usual.” Instead, the administrators felt the spike in negative student behavior may have been due to change in the ODE’s rules for reporting behavior incidents. The Gear Up school deans were not positive, but speculated this year may have been a year where ODE counted Alternative Placements against the school with regards to discipline reporting.

**Discussion**

*Gear Up academic outcomes related to CSR literature*

Academic achievement findings from this study must be interpreted with caution as there were no baseline data for comparison. Since the OATs were first administered in the fourth year of Gear Up CSR program implementation, there was no way of knowing how students would have performed on these tests before and in the early years of Gear Up implementation. This being said, some of the academic findings from this study are relevant to other CSR program findings with regard to teacher cohesiveness, support, and time.

**Teacher cohesiveness and academic outcomes**

Prior research suggests that improved school culture or teacher communities through CSR efforts increase the fidelity of
implementing a school reform program (Graczewski, Ruffin, Shambaugh, & Therriault, 2007; Kurki, Boyle, & Aladjem, 2006). Further, higher levels of CSR program implementation have been associated with higher levels of student achievement outcomes (Berends et al., 2002). Reading OAT score results from the Gear Up school, where scores improved over time, promotes this line of thinking in the literature. To explain, three of the six language arts teachers had been at the Gear Up school from either before the CSR began or within the first two years of its implementation. These teachers helped to establish the fundamentals of what the CSR program would look like from the ground level. Therefore, many of these teachers had a vested interest in the CSR program and teacher clusters. As reported from administrators and teachers alike, the language arts teachers had formed a “tight knit” department where Gear Up helped “them have an improved sense of community.” According to the administrators, “veteran teachers brought new teachers into their clusters by welcoming them and getting them on board with their plans.” Therefore, even though there was some degree of turnover in the language arts department, half of the department remained stable and intact, providing leadership, support, and direction to newer teachers.

Conversely, the math department had experienced massive teacher turnover. One teacher having the most years teaching experience in the Gear Up school of three years—beginning in the 5th year of Gear Up CSR implementation. High levels of teacher turnover is one of the leading reasons CSR programs fail; the new teachers do not buy into the current program and thus do not implement the program with high levels of fidelity (Herrmann, 2006). Therefore, student math OAT scores’ significant increase over time seems somewhat contrary to what current literature would suggest since most of these teachers never had the opportunity to provide input on the school restructuring initiative or to work in a stable community setting for very long. However, according to the administration, the high level of teacher turnover in the math department was necessary to gain better quality mathematics teachers.

Teacher support and academic outcomes

Although there were relatively high levels of teacher turnover in the Gear Up school over the seven years of the CSR program’s implementation, there were vast support networks in place throughout the school to keep the reform efforts strong. Design team support and assistance through professional development and materials are commonly identified factors leading to higher levels of CSR program implementation fidelity (Berends et al., 2002; Datnow & Stringfield, 2000; Graczewski et al., 2007; Herrmann, 2006; Kurki et al., 2006; Vernez & Goldhaber, 2006) and further suggested to increase student achievement outcomes (Berends et al., 2002). During the Gear Up implementation years, Gear Up professors were in the school frequently—2–3 times a week. According to the teachers, the Gear Up professors and faculty ran teacher cluster leader meetings to identify issues of concern, materials needed, or professional development topics of interest. Gear Up student OAT results over time and teacher reports support the notion that outside CSR support, materials, and funding may help to improve student achievement results and also sustain a CSR program initiative.

While external support is known to help increase CSR program implementation and sustainability, lack of funding has been identified as the greatest barrier to sustaining a CSR program (Berends et al., 2002; Tushnet & Harris, 2006; Vernez & Goldhaber, 2006). Teachers felt lack of funding from the CSR program from 2008 to the present did reduce some afterschool program options for students, but did not in any way impact teaching or teacher collaboration learned from the Gear Up reform efforts. Datnow and Stringfield (2000) suggested that “for a reform to lead to meaningful change, it needs to become part of the fabric of a school, not just another passing fad” (p. 196). According to the Gear Up teachers, the instructional changes and school restructuring that occurred as a result of the Gear Up CSR program had become a part of who they are now and was not likely to change. Teachers’ continued commitment to the school restructuring plan (e.g., clustering, behavior plans, curriculum mapping) without monetary or external support reflects this attitude.

Time and academic outcomes

For CSR programs, time could mean two different things. Time could refer to the program’s years of implementation in a school or to the amount of time a student has participated in the CSR program. Limited research on the study of longitudinal CSR program implementation has shown that the longer a CSR program is implemented the greater the student achievement gains (Borman et al., 2002; Zhang et al., 2006). And, greater student achievement gains are made by students who are exposed to the CSR program for longer periods of time (Borman et al., 2006; Herlihy & Kemple, 2004). Quantitative findings from this study support the belief that the longer a CSR program is implemented, the greater the student achievement gains as student achievement results increased to above the proficient level the longer the Gear Up CSR program was implemented. While this trend appears true at surface level, we cannot be certain how much of this trend is due to the CSR program or other factors such as Title I math tutoring indicated by teachers, or teacher turnover in the math department as suggested by administrators. Regarding student exposure to a CSR program and academic achievement, this study was not designed to effectively evaluate this question since it only investigated differences within grade levels over time (comparing same grade cohorts over consecutive years) rather than between grade levels over time (comparing the same students as they move from one grade to the next).

Gear Up non-academic outcomes related to CSR literature

Unfortunately, little research has been conducted specifically examining student behavior and attendance outcomes as a result of CSR programs. Attendance rate results for this study were in line with the limited research which suggests attendance rates improve as a result of CSR programs (Herlihy & Kemple, 2004; Munoz, Ross, & McDonald, 2007). Further, it would not be surprising to see attendance rates increase and behavior incidents decrease in a school running an effective CSR program since the goal of CSR programs is to impact the students, teachers, and school holistically.

Interesting findings from this study revealed that there are multiple problems associated with accurately measuring the impact of a CSR program on non-academic student outcomes. External factors such as policy changes and unforeseeable events may greatly influence these variables and make it challenging to measure actual impact. In this study, the effect from policy changes with regard to attendance rate and behavior incident reporting cannot easily be disentangled from the actual effect the CSR program may have had on these variables. Based on qualitative information obtained from the teachers, it seems likely that student behavior and attendance did improve over the years as a result of Gear Up implementation in their school. However, the extent of the impact Gear Up actually had on these variables cannot truly be determined quantitatively due to the state policy changes where attendance and behavior incident reporting procedures varied across the eight years of this study.

Regardless of how long a CSR program has been implemented, unforeseeable events could significantly impact the school...
environment, teacher attitudes, and student performance. According to the teachers in this study, the unforeseeable event of the 2005–06 mid-school year announcement of the school building closing with no plan for relocation greatly impacted teachers, parents, the school environment, and ultimately students. As such, the “unsettled” and “unpleasant” atmosphere likely led to a significantly large spike in reported student behavior incidents for this year even though the cluster behavior plans had been developed and the CSR program was considered stable in its 5th year of implementation.

Conclusions, implications, and limitations
To be taken seriously and considered scientifically based research by the federal government, rigorous experimental methods are necessary for evaluating the efficacy of educational reform efforts and their effects on student outcomes (Borman et al., 2002; Crowley & Hauser, 2007; Slavin, 2002). While these quantitative methods certainly have earned their place in educational research, it is also important to remember that schools are complex systems where numerous factors are constantly at work and at times acting in opposing directions. Without qualitative explanations of why changes are occurring, results from studies utilizing the most rigorous of research designs may hold little credibility with regard to inferring causation due to numerous confounding variables potentially interacting with the reform.

This study not only reports on the effectiveness of the Gear Up program, but also demonstrates how potential explanations for the quantitative results observed would not have surfaced without the qualitative data. Although there were multiple instances where qualitative information helped to inform and better reveal the quantitative findings in this study, one specific example illustrates how the quantitative results may have been interpreted differently without the qualitative explanations from teachers and administrators. Quantitative findings indicated that math OAT scores significantly increased each year from the 2004–05 school year to the 2006–07 school year moving students from an an average Limited score (lowest rating) to Proficient. Without qualitative information from the teachers and administrators in the school, it would have been easy to speculate that the significant increase in math OAT scores was at least in part due to the fact that Gear Up continued to function and provide monetary support in the school.

CSR research would have supported this interpretation based on the quantitative data alone since it has been shown that the greater the implementation time the greater the student achievement gains (Borman et al., 2002; Zhang et al., 2006). However, when teachers were questioned about this significant increase in math student achievement, they firmly believed that the additional Title I services in the 2007–08 school year are what pushed that group of students to average above proficient. Additionally, administrators suggested that teacher turnover was responsible for the overall increase from year to year, contrary to what current research would suggest. In this instance, qualitative findings helped to inform and clarify the quantitative findings that may have been interpreted in a vastly different way had “insider information” from the teachers and school level administrators not been available or solicited. Noteworthy, however, is that the qualitative data sources reported on in this study were limited to focus group interviews with these stakeholders and program activity reporting documentation. Additional observations, one-to-one interviews, and document analyses would have potentially provided for additional insights into why the quantitative results were observed. This latter point relates to a limitation of sequential explanatory designs in evaluation research in general in that the quantitative and qualitative data are not collected concurrently and greater emphasis is placed on the quantitative data.

Possessing baseline data for attendance rate and behavior incidents allowed for a quasi-experimental cohort design to be implemented for the purpose of drawing causal inferences, as “quasi-comparability” can often be assumed between the cohorts (Cook & Campbell, 1979). However, in this study, the question of “why” these outcomes changed over time would still have been missing without the addition of the qualitative data. Further, attributing the increased student attendance rates and decreased behavior incidents entirely to the Gear Up CSR program would have been inappropriate in this study since there were also dramatic policy changes with regard to how attendance rate and behavior infractions were counted over the years. Researchers must be aware when the statistics they are most interested in studying undergo definitional changes due to state or federal policy. It is a researcher’s responsibility to ensure the statistics they use are either defined similarly for the years of their study or to explain the changes and potential impacts any differences may have on the results. Additionally, the nature of a bottom-up school reform, such as this one, presents a threat to the internal validity of the study in that the treatment (i.e., small school clusters and continuous professional development) slightly shifts from year to year as it is revised based on formative feedback from the evaluation. While this seems like good practice in a practical sense, even slight modifications to the treatment may impact the evaluative ability of the treatment since all students are not experiencing the exact same conditions over time. Without standardized measurement of variables over time, comparisons across years may not be very meaningful (Best, 2001) even with the most rigorous of research designs.

While Gear Up most likely did contribute to positive changes in both academic and non-academic outcomes, as reported by the teachers and administrators, the extent to which Gear Up improved these outcomes is still largely unknown. Therefore, it is important to realize the limitations even experimental studies may have in educational research when further qualitative investigation into the contextual settings of the schools is not done. This being said, the reader must be cautious of the meta-inferences drawn regarding the impact of the CSR program on these outcomes. Caution in making meta-inferences must be taken in any case where non-identical samples are used in the quantitative and qualitative phases (Onwuegbuzie & Johnson, 2006; Teddlie & Tashakkori, 2009). In this case, student data were analyzed in the quantitative phase while teachers and administrators were interviewed in the qualitative phase. It is possible that students might offer a different explanation to the anomalies observed in the quantitative data.

This study explicitly illustrates how using statistics or narrative alone does not always provide a holistic picture of a CSR program’s impact on student outcomes. By combining the two methods in a purposeful way, inferences become more meaningful and policy decisions better informed.

Appendix A. Interview protocol for behavior infractions
- Overall, there was a significant decrease in behavior incidents at your school from the year before Gear Up began (00–01—average 2.09 behavior incidents per student) to its last year of implementation (07–08—average 1.42 behavior incidents per student).
- Why do you think this might have occurred?

- In the 05–06 school year, there was a significant increase in behavior incidents from 04–05 (average 1.61 behavior incidents per student in 04–05 to average 2.27 behavior incidents per student in 05–06).
  - Why do you think this might have happened?
  - What was going on in the school this year?
  - New principal? New discipline policy?
  - What may have impacted student behavior and in what ways?
- In 06–07, behavior incidents again significantly decreased getting back to what was similar for the 04–05 school year.
  - What may have happened to lower the behavior incidents again?
- Did Gear Up interventions specifically target student behavior?
  - Any Gear Up programs designed specifically to help with this area? How?
  - If yes, was there high student participation?


