The Student Achievement Effects of Comprehensive School Reform

A Canadian Case Study

John A. Ross*
Peter Gray
Tim Sibbald
University of Toronto

Abstract

The student achievement effects of Comprehensive School Reform (CSR) are difficult to determine, especially in CSR programs that are based on a set of principles that are enacted by struggling schools into a school improvement process. We examined the effects of a program to support struggling schools in a Canadian province by comparing the 83 low achieving treatment schools to a control group sample of 83 schools matched on SES profile and school district. Scores on the provincial Reading assessment were impressively higher in the treatment than in the comparison group (ES=.70). Statistically significant improvements in school performance were reported as early as two years into the 4-year program and were still significant for the cohort measured one year after the program’s end. These results compare favorably to the effect sizes reported in meta-analyses of CSR (ES=.15 Borman, Hewes, Overman, & Brown, 2003) and Title I (ES=.11 Borman & D’Agostino, 1996) programs.

*Corresponding Author
Dr. John A. Ross, Professor, Curriculum, Teaching, & Learning,
Ontario Institute for Studies in Education,
University of Toronto,
Box 719, 1994 Fisher Dr.,
Peterborough, ON K9J 7A1
CANADA

E-mail: jross@oise.utoronto.ca
Tel: 705-742-9773, Ext. 2206
Fax: 705-742-5104


The authors express their appreciation to David Johnson of Wilfrid Laurier University for providing access to the C.D. Howe database on the SES characteristics of elementary school students.
Comprehensive School Reform (CSR) refers to school improvement efforts that engage all students and teachers in participating schools in improvement efforts, rather than targeting needy individuals. CSR is multi-dimensional, typically changing instructional practice, strengthening professional networks within and among schools, forging relationships with outside agencies, and in some instances building parent capacity. CSR derives its name from the Comprehensive School Reform Demonstration Act of the United States which replaced Title I. Its purpose is to support whole school improvement plans that use research-based practices to improve student achievement (McCaslin, et al., 2006). Although the term CSR emerged in an American context, other industrialized nations have similar programs and frequently refer to their efforts as CSR.

In this article we report a study examining the student achievement effects of a CSR program, Struggling Schools (a pseudonym), implemented in a Canadian province. We will begin by distinguishing the features of CSR that were embedded in Struggling Schools, describe the logic model, and assess the program’s impact on student achievement.

Background

Approaches to Comprehensive School Reform

Two broad approaches to CSR are visible in the literature: (1) specific evidence-based CSR programs developed by external agencies and (2) user generated programs developed from evidence-based principles using a school improvement model. An extensive series of evaluations has examined the effects of the first type, specific CSR programs, based on field trials of well-specified programs that require high fidelity implementation by users. For example, Success for
All has demonstrated consistent achievement effects, especially in K-3 Reading (e.g., Slavin, Borman, Cheung, Chamberlain, Madden, & Chambers, 2006).

Alternatives to specific CSR programs emerged from studies of why a program that was highly successful in one site failed in another. Hatch (2000) found that lower achieving schools lack the capacity to select the specific CSR program most suited for their context or to implement the program as intended. Other researchers found that the specific CSR program chosen by schools might align poorly with other structures. For example, Greene and Lee (2006) found that a strong instructional program (Problem Based Learning) was not implemented by teachers because they believed it conflicted with preparation for state assessments. Datnow (2001), in a study of the sustainability of CSR, found that conflict with state assessments was a factor in all the schools in her study that dropped CSR. A further limitation of these specific CSR programs is that they may deskill teachers (Hargreaves, 1994).

The second type consists of a school improvement process based on CSR principles. The Struggling Schools program was of this kind. In this approach schools do not implement a specific program but design, usually in consultation with outside agencies, a school improvement plan that incorporates principles of school change associated with higher student achievement. Typical steps in the process include conducting an audit of the school’s strengths and weaknesses, setting priorities and targets in a set of actions, executing the actions, and evaluating outcomes.
Program Description

The purpose of Struggling Schools was to increase K-3 Reading in low achieving schools. There was staggered entry into the program: in 2001-02, 15 schools were admitted; in each of 2002-03 and 2003-04 there were 14; and in the final year 2004-05 there were 57 schools; i.e., \( N=100 \) schools in the program. Each school was admitted for four years: three years of intensive support followed by a fourth exit year of reduced support.

Schools and their districts applied for admission to the program. To be eligible, the school had to have had less than a third of its students meeting provincial standards for grade 3 and 6 in the annual assessments conducted by the province’s external testing agency. In addition, the principal of the school had to be willing to lead an improvement effort and the district had to be committed to supporting the school over the four years of the program (e.g., by agreeing not to move the principal during this period—Hatch, 2000 and Yap, 2005 found leadership turnover to be an impediment to CSR implementation). The CEO of each applying district served on the selection committee and provided the information used by the Steering Committee to choose schools. Treatment schools reflected provincial variation in district characteristics: language, religion, region, and student demographics (i.e., majority, minority, and aboriginal cultures).

Once in the program, schools developed an inventory of their literacy resources and formed a school improvement team. A diagnostician (typically university-based) who was believed to be an expert in literacy instruction and school change visited the school to assess its strengths and weaknesses. The diagnostician delivered a report that identified priorities for change and recommended actions to the school. Principal and teaching staff designed a school improvement plan based on the diagnostician’s report. After provincial approval of the plan, the school received financial resources (e.g., for release time for teacher in-service and purchase of
professional learning materials) to implement the plan. A case manager assigned to the school delivered or coordinated teacher training on literacy skills (e.g., on differentiated instruction), while a leadership advisor advised the principal on strategic planning (e.g., techniques for creating professional community in the school). At the end of the year the diagnostician returned for observation and feedback on the school’s progress.

The second and third years followed the same format. The fourth year focused on sustainability. The school conducted its own review of progress and generated an exit report with procedures for continuous improvement. The final diagnostician report provided feedback on plans for school improvement beyond the funding period.

Program Theory

The program theory of Struggling Schools stated that the problem to be solved was persistent low achievement on mandated assessments. The causal mechanisms to be activated by the intervention were based on Fullan’s (2002; 2005) theory of action which highlights a small set of principles and strategies. The most important for the Struggling Schools program were capacity building, partnership, and accountability (Fullan & Campbell, 2007).

Capacity building “consists of developments that increase the collective power in the school in terms of new knowledge and competencies, increased motivation to engage in improvement actions, and additional resources (time, money, and access to expertise).” (Fullan, 2005, p. 175) Research on school change suggests that capacity building has at least three dimensions. The first is the acquisition of research-based technical skills to improve student learning. For example, a randomized field trial found that teachers who learned how to use a technical skill, Running Records assessment to target reading instruction, had higher student achievement than control group teachers (Authors, 2004). The second dimension of capacity
building is the creation and maintenance of a supportive organizational structure. Louis and Marks (1998) found that professional community (consisting of shared values, a focus on student learning, collaboration, commitment to deprivatized practice and reflective dialogue) was a strong predictor of social support for achievement, quality of classroom instruction, and authentic student achievement. Technical skill development and the creation of a supportive organizational structure can be developed simultaneously if professional development focuses on collective teacher learning (Watson, Fullan, & Kilcher, 2000). The third dimension of capacity building is leadership, particularly leadership that raises the values of members, motivating them to go beyond self-interest to embrace organizational goals, and redefining their needs to align with organizational preferences. Authors (2006) found that school leadership contributed to student achievement indirectly, i.e., through the principal’s enhancement of collective teacher efficacy and strengthening of teacher commitment to the mission of the school.

Capacity building in Struggling Schools consisted of in-service on the technical skills of literacy instruction as determined by the report of the diagnostician, which was based on the province’s Expert Panel on Early Reading. The Expert Panel reviewed research on reading to prescribe specific instructional and assessment strategies to improve student performance. The case manager visited each school monthly to coordinate in-service (drawing upon the resources of local experts), advised the principal about how to enhance the capacities of his/her staff, and provided peer coaching and mentoring to individual teachers about how to meet the reading needs of particular students. In-service was delivered within professional communities in order to simultaneously develop individual and collective capacity. Principal capacity building focused on the creation of professional communities in the school, following the prescriptions of the Expert Panel report. The case manager met for half a day monthly with the principal and the
leadership advisor met with the principal for a full day every six weeks, more frequently for isolated principals in remote locations. Project funds provided release time for teachers and the purchase of materials for professional learning.

**Partnership** refers to the development of productive relationships among schools (lateral capacity building) and between schools and external agencies (vertical capacity building). By interacting with local and external experts, schools can strengthen their ability to move forward. The meta-analysis of CSR initiatives by Rowan, Barnes, and Camburn (2004) found that CSR worked best when external agents produced a high quality school improvement design and supported teachers as they implemented it. Yap (2005) found there were positive effects when external agents provided technical assistance on specific school reform processes to the principal and staff in underperforming schools. Support for underachieving schools can be provided by LEA consultants (Chapman & Harris, 2004), a case manager assigned by a national project (Chapman & Allen, 2006) or school district agents (Stoll, Bolam, McMahon, Wallace, & Thomas, 2006). In terms of local school networks, Wohlstetter, Malloy, Chau, and Polhemus (2003) found that networks of high school and feeder schools enhanced organizational capacity when the schools were actively engaged in reform, a theme also supported by other studies (Chapman & Allen, 2006; Kahne, O’Brien, Brown, Quinn, 2001).

Partnerships in Struggling Schools emphasized building links between schools and external agents with little attention to forging networks of schools within the program or in the schools’ districts. The diagnostician, the case manager, and the leadership advisor provided schools with knowledge and skill they lacked on how to assess their needs, monitor their progress, and improve teaching and learning. The school improvement plan, which required
provincial approval, was the blueprint for their intervention. The agents, who were coordinated by provincial managers, developed capacities of teachers and/or principals.

*Accountability* refers to the process of setting school targets, measuring performance, and identifying data-based, ameliorative strategies. Yap’s (2005) study of underperforming schools found that external agents played a key role by diagnosing school needs. Chapman and Harris (2004), among many researchers, emphasized the importance of creating a data rich environment in which assessment data guides instruction (cf., reviews by Muijs, Harris, Chapman, Stoll, & Russ, 2004 and Black & Wiliam, 1998). In addition, some researchers found that external accountability systems motivate school improvement and lead to higher achievement. For example, Hanushek and Raymond (2005) found that “states that introduced consequential accountability systems early displayed more rapid gains in NAEP [National Assessment of Educational Progress] performance, holding other inputs and policies constant” (p. 310), a claim also supported by the NAEP analysis of Carnoy and Loeb (2002). However, negative effects of accountability systems have been frequently noted. For example, Booher-Jennings (2005) found that teachers preparing for state assessments practiced educational triage; i.e., they allocated instructional resources to students who were close to passing and ignored those perceived to be hopeless and the easy passers. Earl and Katz (2003) argued that the challenge is to move from using data as a surveillance activity to using it in the service of improvement, a theme also emphasized by Fullan and Campbell (2007) who argued that focusing on accountability prior to building capacity impedes the development of partnerships. Fullan (2006) emphasized the need to integrate internal and external accountability systems, as illustrated by the case studies in Wilson (2004).
In Struggling Schools, accountability was invoked through school level and individual student level data. School level data consisted of external provincial assessments. These annual reports reported the proportion of students who met the provincial standard; increasing this proportion was the ultimate criterion for measuring school success. In addition, the year end diagnostic visit focused on changes from the initial visit in terms of the instructional practices recommended by the Expert Panel report. Student level accountability consisted of encouraging teachers to use multiple measures of students’ reading performance to plan instruction. Student data, such as the number of children reaching reading levels defined by programs like First Steps (Department of Education, Science, and Training, 2003) were displayed in teacher staff rooms.

**Research Questions**

The province sought an external agency to assess Struggling Schools because it wanted to avoid self-interest bias. Borman et al. (2003) found that CSR programs reviewed by developers generated significantly higher effect sizes than programs reviewed by third parties. Our research questions were: (1) What were the effects of Struggling Schools on grade 3 Reading achievement? (2) Were the effects on Reading achievement moderated by the number of years a school was in the program?

**Quantitative Challenges in Assessing CSR**

**Lack of Boundaries** The school improvement approach to CSR lacks the tight boundaries that characterize the specific CSR program approach. The three principles outlined above and others that have been proposed (e.g., Leithwood, Jantzi, McElheron-Hopkins, 2006) can be enacted in many ways, making it to difficult to aggregate findings. In addition, the school improvement planning approach to CSG swallows up previous school initiatives, making it difficult to separate new dimensions from earlier initiatives (Yin & Davis, 2007).
Unrealistic Expectations. It is not unusual for practitioners to expect that a CSR program will raise a school’s student achievement from the 10\textsuperscript{th} to the 30\textsuperscript{th} or even the 50\textsuperscript{th} percentile. Hedges (2002) challenged these expectations by using NAEP data to show that such increases imply effect sizes of .50 and 1.28 respectively. Hedges described the latter as a monumental achievement, were it to occur. But meta-analyses of comparable programs report considerably lower impacts. Borman et al. (2003) found a mean effect size for CSR programs of only ES=.15, comparable to the ES=.11 reported in a meta-analysis of Title I programs (Borman & D’Agostino, 1996).

Insufficient Statistical Power. Statistical power, the ability of a research design to detect a treatment effect, can be a formidable challenge in assessing the effects of CSR. Statistical power is a function of the number of units in the study, as well as other design features such as the alpha level. Because the whole school is involved in CSR, the number of treatment cases is typically small. Even large studies rarely have as many as sixty schools. Bloom, Richburg-Hayes and Black (2007) demonstrated that if sixty elementary schools were assigned to treatment and control groups, the detectable effect sizes ranged from ES=.27 to .35, considerably higher than the mean effects reported for comparable programs of ES=.11-.15. Impacts that were any lower would lead researchers in a typical study to conclude that the CSR program had no significant effect on achievement. Yet, if more schools had been involved, the study might well have detected a statistically significant achievement benefit.

Bloom et al. (2007) suggested several strategies to raise the statistical power of CSR studies. These strategies include matching schools on a criterion (such as prior achievement or student characteristics) prior to the assignment to treatment and control conditions. Even more powerful, is the inclusion of pretest information or other information about study participants as
covariates in the analysis. Bloom et al. found that using covariates was more powerful than matching strategies but each had an independent effect on power, provided that the matching variable or covariate correlated with achievement. Researchers would be wise to use both strategies. Bloom et al. also reported three findings that were of particular interest to the design used in our study: They found that school level pretests were almost as effective as student level pretests in enhancing power, an important finding because school level pretests are easier to access than student level information. Bloom et al. also found that the contribution of pretests to statistical power declined only slightly as the number of years from pretest to posttest increased. Finally, Bloom et al. found that increasing the number of years in the pretest baseline had little effect beyond using a single year. However, even when these precision enhancing strategies are embedded in the design, detecting the typically small effects of CSR continues to be a challenge.

**Inability to Invoke Experimental Designs.** Randomized field trials provide the strongest protection from threats to internal validity (Cook & Payne, 2002) yet they are rarely used in educational policy evaluation (Burtless, 2002). Studies of specific CSR programs frequently allow participants to decide which CSR innovation they wish to participate in, school improvement approaches rarely provide for control groups, and evaluations of CSR are often post hoc rather than a priori. However, Boruch (2007), while defending the true experiment as the optimal design, noted that quasi-experimental designs can be justified if pretest and other pre-intervention measures are used as covariates.

**Evaluation Timing.** Borman et al (2003) found that the best results were obtained after a school had been in the CSR program for five years but most programs were evaluated after three. Cook, Murphy, and Hunt (2000) found that when Comer’s CSR program was evaluated after two years there were negligible effects but after four years significant achievement effects were
detected. Our knowledge of the enduring effects of CSR is limited because there is little research on program outcomes beyond the funding period (Coburn, 2003).

These challenges to quantitative assessment of CSR affected our research in several ways. First, we recognized that our ability to detect effects of Struggling Schools would be limited by our sample size. There were only 100 schools in the program, admitted over a four year period in cohorts of unequal size. We felt we would be able to detect an effect for all schools but we doubted our ability to find differences between cohorts. Second, we recognized that we needed to take steps to maximize the power of our limited sample by matching treatment to control schools on student background characteristics that correlated with achievement and by using school pretest data and student background as covariates in the analysis. Third, we recognized that although the evaluation was commissioned after all four cohorts had begun, a quasi-experimental design with appropriate controls could address claims of causality. Fourth, the inattention to the enduring effects of CSR persuaded us to observe program effects after the funding period ended.

Method

The design of the study was a quasi-experimental, pre-post matched sample design.

Sample

For each of the 100 schools participating in Struggling Schools we selected a control school that was not involved in the program. We matched on two criteria: school district and composite SES (socio-economic status) score, both of which are correlated with student achievement.

We calculated an SES score for each of the 4,054 schools in the census database of provincial elementary schools compiled by the C. D. Howe Institute. Johnson (2005) reported a
table describing the relationship between socio-economic school communities and grade 3 provincial assessment scores. Using the beta weights in Johnson’s table, we developed a formula for calculating the SES scores of each elementary school in the province:

\[
SES = -0.11X_1 - 0.04X_2 - 0.50X_3 + 0.08X_4 + 0.02X_5 + 0.05X_6 + 7.20X_7 - 0.40 X_8 - \\
0.08X_9 - 0.23X_{10} - 0.20X_{11} + 0.28X_{12} + 0.005X_{13} - 0.000003X_{14}
\]

Where \( X \) represents the following variables in the database:

\( X_1 \) Percent of families with children that are single parent families
\( X_2 \) Percent with English or French as home language
\( X_3 \) Percent aboriginal in census
\( X_4 \) Percent recent immigrants (in Canada less than 5 years)
\( X_5 \) Percent born outside Canada, U.S and U.K
\( X_6 \) Percent of dwellings that are detached homes
\( X_7 \) Log of Average household income 1995$
\( X_8 \) Percent of households that moved in the last year
\( X_9 \) Percent of households that moved in the last 5 years
\( X_{10} \) Unemployment rate
\( X_{11} \) Percent without a secondary school diploma
\( X_{12} \) Percent with some university education
\( X_{13} \) Total enrolment - average over 3 years of postal code data
\( X_{14} \) Total enrolment squared

Within each district, we selected as the matching control the school with the SES value that was closest to that of the school participating in Struggling Schools.
Sources of Data

Grade 3 Reading scores came from criterion referenced assessments conducted by an assessment agency independent of the provincial government. The agency produced detailed tables showing how its items mapped on to specific curriculum standards—the items themselves were not published. Provincial curriculum standards specified levels of proficiency (several categories denoting performance below level one and levels one through four). Curriculum documents defined what each level meant in general and in the context of specific grades, subjects, and type of learning objective. Reading scores were based on 32 multiple choice and 12 open ended items to produce a global score. The key measure was the proportion of students in each school who achieved the provincial standard (level 3 or 4) in each year. We used the global score, exempting no students.

The assessment agency used three procedures to ensure reliability: 1) group marking: during training, all markers scored the same student and the results were photocopied for discussion; 2) reinsertion, i.e., a sample of papers was scored by two or more markers; 3) if a marker was in the top or bottom 5% for levels awarded on a given day, that person's output was remarked to guard against leniency/severity differences. The assessment agency does not report reliability coefficients. However, in the 2003 administration, a reliability study found that markers did not substantially influence scores: generalizability coefficients averaged .76 (Dunn, Childs, Cleland, Pang, & Saunders, 2004). Year to year equating is done with four test booklets (each student receives one). Each year one test booklet is retired and another is produced (Wolfe, Childs, & Elgie, 2004). We downloaded school data from the assessment agency’s website. Since the agency suppresses results from schools with less than 15 students in a grade, our sample reduced to 83 school pairs.
Analysis

We defined prior achievement as the mean score in the three years prior to that school’s entry into the program. However, differences in reporting provincial assessment results meant that prior achievement was defined differently for Phases 1 and 2 than for Phases 3 and 4, as shown in Table 1. With this database we conducted GLM (General Linear Modeling) analysis in which the dependent variable was grade 3 2005-06 Reading scores. The covariates were prior achievement in Grade 3 Reading and school SES. The independent variables were group (treatment or control) and years in the program. We repeated the analysis using Grade 3 Reading scores for 2005-06 as the dependent variable.

Table 1

Database for Prior Achievement Scores

<table>
<thead>
<tr>
<th>Years in Program</th>
<th>Phase</th>
<th>Year of Entry</th>
<th>Source of Data for Prior Student Achievement</th>
</tr>
</thead>
<tbody>
<tr>
<td>One year from exit</td>
<td>1</td>
<td>01-02</td>
<td>00-01</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>02-03</td>
<td>00-01,01-02</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>03-04</td>
<td>00-01,01-02,02-03</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>04-05</td>
<td>01-02,02-03,03-04</td>
</tr>
</tbody>
</table>

This design had high statistical power for determining whether there was an overall effect of the Struggling Schools program, i.e., when the effects for all phases were aggregated. With a sample size of 166, a change-score control design and equal cell sizes, we were able to detect a program effect as small as ES=.17 with 80% power at p<.05 (Dennis, 1994). However, comparisons between cohorts were underpowered because of small cell sizes.

Results

There were no significant differences between the treatment [mean=10.06, SD=5.74] and control [mean=11.46, SD=4.79] groups on school SES [t(164)=-1.699, p=.091]. However, prior
reading achievement was significantly lower for the treatment [mean=.26, SD=.09] than the control [mean=.41, SD=.09] schools \([t(130.289)=-7.67, p<.001]\).

Table 2 displays the covariate adjusted Grade 3 Reading means and standard deviations for each of the four cohorts and their matching controls in 2005-06. The GLM analysis explained 18% of the variance in 2005-06 achievement. There was a significant program effect \([F(1,156)=11.525, p=.001]\) accounting for 6.9% of the variance, which was smaller than the effects of prior achievement \([F(1,156)=16.329, p<.001, eta^2=.095]\) but larger than the effects of SES profile which was not statistically significant \([F(1,156)=3.768, p=.054, eta^2=.024]\). Schools that participated in Struggling Schools had higher Reading achievement than schools that did not. The effect size across all schools, based on the covariate adjusted means and weighted by number of schools, was .70. In comparing prior achievement to the 2005-06 results, we noted that the proportion of students meeting the provincial achievement standard in Reading improved in both treatment and control schools but the gains were four times as large in the treatment (107% increase) than in the control schools (27% increase).

There was no group * years effect. We found that Struggling Schools had a statistically significant effect in three of the four comparisons: after two years in the program \([t(95)=2.016, p=.023, ES=.37]\), after three years \([t(18)=2.206, p=.020, ES=.95]\), and one year after exit from the program \([t(23)=2.539, p=.012, ES=.70]\). The adjusted achievement mean was higher in the treatment than in the control after four years in the program but the difference was not statistically significant \([t(19)=1.319, p=.10, ES=.54]\) due to the small number of schools.
Table 2
Covariate Adjusted 2005-06 Grade 3 Reading Means and Standard Errors by Study Condition and Number of Years in the Program

<table>
<thead>
<tr>
<th>Group</th>
<th>Years</th>
<th>Mean</th>
<th>Standard Error</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treat</td>
<td>2</td>
<td>.55</td>
<td>.02</td>
<td>49</td>
</tr>
<tr>
<td>Control</td>
<td>2</td>
<td>.48</td>
<td>.03</td>
<td>49</td>
</tr>
<tr>
<td>Treat</td>
<td>3</td>
<td>.64</td>
<td>.05</td>
<td>10</td>
</tr>
<tr>
<td>Control</td>
<td>3</td>
<td>.50</td>
<td>.05</td>
<td>10</td>
</tr>
<tr>
<td>Treat</td>
<td>4</td>
<td>.56</td>
<td>.05</td>
<td>11</td>
</tr>
<tr>
<td>Control</td>
<td>4</td>
<td>.46</td>
<td>.05</td>
<td>11</td>
</tr>
<tr>
<td>Treat</td>
<td>One year from exit</td>
<td>.64</td>
<td>.05</td>
<td>13</td>
</tr>
<tr>
<td>Control</td>
<td>One year from exit</td>
<td>.51</td>
<td>.05</td>
<td>13</td>
</tr>
<tr>
<td>Treat</td>
<td>All</td>
<td>.48</td>
<td>.03</td>
<td>83</td>
</tr>
<tr>
<td>Control</td>
<td>All</td>
<td>.40</td>
<td>.04</td>
<td>83</td>
</tr>
</tbody>
</table>

Discussion

What were the effects of the Struggling Schools program on grade 3 Reading achievement?

The impact of the program was substantial. The effect size for all schools was .70. These effect sizes, especially the latter, compare favourably to the ES=.15 for all CSR programs in the meta-analysis of Borman et al. (2003) and ES=.11 for all Title I programs (Borman & D’Agostino, 1996).

As noted earlier, attributing the effects of CSR to specific program elements is difficult because principle based CSR programs are under specified. Determining where CSR begins and ends and how it differs from previous reform efforts in a particular site is notoriously difficult. We attributed the effects on student achievement to the three principles implicit in the program design:
First, Struggling Schools provided substantial resources to increase the capacity of under-achieving schools. The program funded release time for teachers to attend in-service on literacy skills and professional learning materials to enable teachers to expand their technical skills. The monthly visits by staff ensured that the in-service would be site-based and focused on specific needs identified in the diagnostic reports. The program enhanced professional community by providing release time for teachers to share their learning experiences with peers and by delivering in-service to professional communities rather than individuals. The program strengthened principal leadership skills by providing regular visits by a leadership advisor. A key feature of these capacity development efforts was the tight alignment of diagnosis, school improvement plan, and capacity development initiatives.

Second, Struggling Schools provided additional capacity to treatment schools through an external partnership. A key process was the external diagnosis, based on the Expert Panel Report on Literacy, which became the foundation for the school improvement plan, the focus of in-service efforts, and the framework for assessing school progress on teaching. However, control of resources was shared. After provincial staff had approved the budget, allocation to specific priorities was determined at the school level.

Third, the program strengthened accountability mechanisms in each school. In addition, to the annual provincial testing, Struggling Schools provided schools with the means to develop other measures of student learning, provided resources for training in their use, and modeled how to use student assessment data to target instruction. The program also provided models of target setting and tracking mechanisms to recognize gains and focus efforts on specific student needs.

The integration of these three elements mitigated the threat of external assessment to teacher beliefs about their capacity. Teachers with high self-efficacy, i.e., teachers who expect
that they will be able to bring about student achievement, set higher goals for themselves and their students, persist through obstacles, are more likely to implement challenging instructional methods, and have higher student achievement (Authors, 1995). Bringing in external diagnosticians could have been interpreted by staff to mean that the province believed they were incapable of diagnosing their school’s needs. Usher and Pajares (2005) found that messages communicating to recipients that they are not competent to complete core tasks reduce self-efficacy beliefs; i.e., the message that staff are unable to diagnose their school’s needs could have become a self-fulfilling prophecy. But this negative message was greatly mitigated by positive messages that affirmed staff competence: inviting schools to develop the school plan with minimal external intervention, allowing schools to allocate project resources within the school improvement plan, building local expertise into in-service efforts, using internal measures to track progress to supplement provincial assessments, and making the last year of the program a transition period in which schools consolidated their capacity and shifted from external to self-direction.

Were the effects on Reading achievement moderated by the number of years a school was in the program?

There were no statistically significant differences among the four cohort groups. The effect sizes varied in no discernable pattern. Using Cohen’s (1988) categories, there was a small effect (ES=.37) after two years in the program, a large effect (ES=.95) after three years, a medium effect (ES=.54) at the end of the fourth or exit year, and a medium effect (ES=.70) one year after exit from the program. The finding that there were no significant differences among the cohort groups suggests, but does not definitively demonstrate, that the cohort groups were
equivalent, i.e., that the criteria for entry were relatively stable over the four years of the program.

**Threats to Internal Validity**

The key limitation of the study is that we cannot be sure that the groups were equivalent prior to the start of the program. We mitigated the effects of using a quasi-experimental design when we matched the schools on two key variables (school district and SES profile) that predict provincial assessment scores and we used prior achievement and school SES profile (both predictors of outcomes) as covariates in the analysis. But without random assignment to study conditions, it is possible that the treatment schools differed from the control schools in ways that we did not measure. The most likely alternate explanation is that “readiness for change”, represented by a school’s willingness to be in the program and its selection by the district’s CEO as having sufficient capacity to benefit from the program, may have contributed to the outcomes. A stronger design would have identified, four years prior to the start of this study, schools eligible for the program and randomly assigned them to treatment and control conditions.

Another limitation of the study is that the covariate (prior achievement) represents the performance of earlier cohorts from the same school—these are not the same children that participated in the 2005-06 assessments. Scores on mandated assessments can vary dramatically from one year to the next (Linn & Haug, 2002). However, these year to year fluctuations within schools wash out as the sample grows larger. With 83 schools (more than 2700 students tested) in each group, we can assume that cohorts were virtually equivalent.

A third limitation is that the number of years on which the prior achievement measured was based varied from one to three years due to changes in how the assessment agency reported
the data. We note, however, that Bloom et al. (2007) found that increasing the number of years in
the pretest baseline had little effect beyond using a single year.

A fourth limitation is that we lost 17% of the treatment schools due to suppressed data.

Directions for Future Research

Evaluations of CSR have focused almost exclusively on determining whether the
program had an effect on valued student outcomes. Very little attention has been paid to whether
these programs are cost-effective. An important exception is Borman and Hewes (2002) who
compared the cost-effectiveness of a specific CSR program (Success For All) to three other CSR
programs. We propose as the next step in our research to examine the cost-effectiveness of
Struggling Schools. Specifically, we propose to compare the cost per child of achieving the
provincial standard in the treatment schools to the cost in the control schools.

Since our evaluation of Struggling Schools, the program has been restructured and scaled
up to involve 800 schools. The new program targets two groups of low performing schools: those
with less than 34% of students meeting the provincial criterion in Reading in any two of the
previous three years and schools with static or declining achievement who are in the range of 34-
50% reaching the provincial criterion. The new program will also provide capacity development
activities for participating districts. A key question is whether the effect sizes for the original
Struggling Schools program will be sustained in the scaled up version.
Conclusion

Quantitative assessment of the impact of type 2 CSR programs, i.e., those based on research principles rather than externally developed programs, is fraught with difficulty. Type 2 CSR programs lack clear boundaries making it difficult to isolate the effects of CSR from other programs operating in the school; effect sizes of CSR tend to be smaller than the typical sample size requires; evaluations of CSR are often post hoc, making it impossible to invoke experimental designs; and programs are often evaluated too early for effects to have accumulated. In this study we found that a CSR program organized around capacity building, external partnerships, and accountability contributed to student achievement. The study demonstrates that principle based CSR programs, as represented by the Struggling Schools initiative, can substantially improve the achievement of students in under-achieving schools.
References


Muijs, D., Harris, A., Chapman, C., Stoll, L., & Russ, J. (2004). Improving schools in socio-


