Cost-Effectiveness of Comprehensive School Reform

in Low Achieving Schools

John A. Ross*
Garth Scott
Tim Sibbald

University of Toronto

Paper presented at the annual meeting of the American Evaluation Association

Conference, Orlando FL, November, 2009.

*corresponding author
Dr. John A. Ross, tel: 705-742-9773 ext 2293
Professor Emeritus & Field Centre Head, fax: 705-742-5104
OISE/UT Trent Valley Centre, www.oise.utoronto.ca/field-centres/tvc.htm
Box 7190, e-mail: jross@oise.utoronto.ca
1994 Fisher Dr., Peterborough, ON K9J 7A1

Abstract

We evaluated the cost-effectiveness of Struggling Schools, a user generated approach to Comprehensive School Reform implemented in 100 low achieving schools serving a disadvantaged student population in a Canadian province. Struggling Schools had a statistically significant positive effect on Grade 3 Reading achievement; *d*=.48 in 2005-06 and .60 in 2006-07. The program was not cost-effective when compared to two alternatives using alternative decision rules. (1) The cost of bringing one student to the provincial achievement standard was more than 25% higher in Struggling Schools than in the status quo (2) The cost-effectiveness ratio (effect size per $1000 of incremental cost) was lower in Struggling Schools than in Success For All. Struggling Schools would have been deemed to be cost-effective if different choices in had been made, especially in the calculation of costs (e.g., the inclusion of donated time), decision rules for declaring cost-effectiveness, and the studies used to access comparative data.

Cost studies in education provide guidance for program improvement by providing and validating models of optimal resource allocation. They encourage implementation of low-cost, moderate-impact programs over high effect-size initiatives that may not be feasible on a broad scale or that lead to lower net benefit for a given budget (Levin, Glass, & Meister, 1987). Cost studies can demonstrate that educational spending provides a substantial return on investment for individuals and society (as in the Perry Preschool studies, Schweinhart, Montie, Xiang, Barnett, Belfield, & Nores, 2005), thereby strengthening public confidence in educational policy making and justifying maintenance of educational budgets. Cost-effectiveness analysis is a particular type of cost study, defined as "the systematic approach of integrating information on the costs and effects of various alternatives to identify the option that most efficiently utilizes limited resources to produce a particular outcome or set of outcomes" (King Rice, 1997, p. 309). Cost-effectiveness studies are particularly helpful when assessing program benefits (e.g., improvements in students' ability to read) that are not easily converted to monetary outcomes, such as career earnings or reduced welfare costs, a requirement of cost-benefit analysis.

In this article we examine the cost-effectiveness of an approach to Comprehensive School Reform (CSR) that was implemented in a Canadian province. Ross, Alberg, and Nunnery (1998) distinguished two approaches to CSR: one in which schools select from a menu of programs developed by external agencies and another in which schools develop a school improvement process using research-based principles with the support of an external agency. Struggling Schools was of the second type. In this article we will briefly review the program (reported in detail in Authors, 2009) and the evidence for claims about its effects on student achievement. But our main focus here is the value for money question: was the Struggling Schools program,

and by extension similar programs launched by states and districts in other jurisdictions, cost-effective? We will present data on the program's costs in relation to its benefits and explore alternate ways of interpreting its cost-effectiveness.

The Struggling Schools Program

The purpose of Struggling Schools was to increase Grade 1-3 Reading in low achieving schools; i.e., those in which less than one third of its students were meeting the provincial standard. There was staggered entry to the program in four Phases: 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. Each school was admitted for four years: three years of intensive support followed by an exit year in which schools transitioned to self-support. Program actions consisted of (i) the school developed an inventory of its resources; (ii) an external diagnostician with expertise in literacy instruction and school change assessed the school's needs and prescribed remedies; (iii) school administrators and faculty developed a school improvement plan; (iv) the province provided funding, tied to the plan, for in-service, release time and professional learning materials; (v) a provincial case manager delivered or coordinated training on literacy teaching skills and a leadership advisor counseled the principal on strategic planning; (vi) the school implemented its plan and (vii) received feedback from the diagnostician at the end of each year on its progress. Each element of the Struggling Schools program was derived from Fullan's (2002; 2005) theory of change. The causal mechanisms of Fullan's theory most relevant to Struggling Schools were capacity building (the acquisition of research-based teaching skills such as the Expert Panel Report on Literacy, the creation and maintenance of a supportive organization, and transformational leadership), partnerships with external agencies, and

accountability (setting school targets, measuring performance and identifying ameliorative strategies) (Fullan & Campbell, 2007).

We conducted a third party study of the student achievement effects of Struggling Schools which was a quasi-experimental, pre-post matched sample design with school as unit of analysis, drawing on two years of achievement data from standardized external assessments. Struggling Schools had a statistically significant positive effect on Grade 3 Reading achievement; *d*=.48 in 2005-06 and .60 in 2006-07, effect sizes larger than those typically reported for well-structured CSR programs. There were no statistically significant differences attributable to year of program entry but there was evidence of enduring achievement effects two years after exit from the program (Authors, 2009).

Determining Cost-Effectiveness

Cost-effectiveness researchers compare the cost-outcome ratios of alternative ways of allocating resources. The practice has been to compare highly diverse interventions so long as they share similar objectives. For example, Levin (2009) examined the cost of five programs (Perry Preschool, First Things First, class size reduction, Chicago child-parent centers, and a 10% increase in teacher salaries) chosen because they were the only interventions for which there is rigorous evidence that the program reduced school dropouts. Educational economists are willing to assume that differences among studies, such as how the outcome variable is measured, the time period of the original data collection, student samples and populations, the scale of the intervention and the mechanisms of its operation, can be measured and controlled within the cost-effectiveness analysis. Harris (2009) viewed the assumption as questionable. He suggested that variability across studies of interventions could be resolved if researchers compared near program substitutes and created tables comparing interventions with similar values on each

dimension of substititutability. Harris's proposal is very much a long term strategy. Because cost studies are infrequently conducted in education (Author, 2007; Hummel-Rossi & Ashdown, 2002; Levin, 2001; Levin & McEwan, 2001) there are few cost studies to draw upon for comparison.

We pushed Harris's suggestion further by attempting to control for some of the most important contextual variables that influence a program's impact. We compared the Struggling Schools program to its nearest equivalent, schools that were highly similar to those receiving the intervention in terms of student characteristics, policy context, and funding, but differed in terms of the specific features of the intervention. In other words, we compared the cost-effectiveness of the program to its most available alternate, the status quo. Although some economists define policy alternatives to exclude the status quo, the organizational literature treats the status quo as a credible policy option (Boyle, DuBose, Ellingson, Guinn, & McCurdy 2001). The schools to which Struggling Schools were compared constituted a near-treatment group. They were teaching Reading in Grades 1-3, attempting to implement the same instructional strategies described in the Expert Panel Report that were the focus of the Struggling Schools; they experienced the same accountability pressures in response to similarly low achievement.

Cost-Effectiveness Comparison to Matched Control Schools

To determine whether Struggling Schools was cost-effective, our first thought was to calculate the average PPE (Per Pupil Expenditure) required to bring one student in the Struggling Schools program to the provincial achievement standard. We could then compare the PPE cost of success in Struggling Schools to the PPE cost of success in similar schools not participating in the program.

This strategy assumed that the cost of increasing the number of students reaching the provincial standard is constant across all student performance levels, an assumption that is unlikely to be valid. Students who are advantaged by high prior achievement, positive dispositions toward learning and ample social capital are easier to teach than students who lack these advantages. It is probable that greater instructional resources would be required to bring underachievers to the provincial standard than is the case for students who are currently meeting that standard. A simple comparison of the PPE costs of successful students to status quo costs would be biased in favour of the status quo. We need to adjust the comparison to recognize that increasing the achievement of unsuccessful students to the level reached by successful students increases unit costs.

Research on mastery learning suggests that the adjustment should be quite large. For example, Gettinger (1985) found that some students reached mastery on a reading task after a single trial; others needed 2-6 trials before they were successful and some were unsuccessful even after six trials. Mastery learning researchers found that even when 5-10% of the lowest achievers are excluded, raising the performance of the less able requires more time than teaching able learners. Arlin (1984) found that students who needed remediation required 36-99% more time to achieve mastery than students who were successful on the first trial. Arlin compared the time required by the slowest 20% of students in several grades to the fastest 20% in the same grades: the slowest grade 3 students required 2-5 times as much instructional time as the fastest. Arlin and Webster (1983) found that after eliminating 15% of students who failed to master the task, the slowest quartile of the remainder took more than twice as much time as the fastest quartile to reach mastery. The three studies in the meta-analysis of mastery learning programs by

Kulik, Kulik, and Bangert-Drowns (1990), which did not include the Arlin or Gettinger studies, also found that mastery learning required additional time, as did Martinez and Martinez (1999).

The findings from mastery learning research demonstrate the need to adjust the comparison of the cost per successful student in Struggling Schools to the cost of success in control schools and provide some guidance about the order of magnitude. The analogy to mastery learning is imperfect but an adjustment of 25% would not be unreasonable; i.e., the Struggling Schools program would be cost-effective if (i) it increased the number of students reaching the provincial achievement standard, and (ii) did so at a cost per successful student of no more than 125% of the cost per successful student of similar students in control schools.

Cost-Effectiveness Comparison to Success For All

We located three studies that examined the cost-effectiveness of CSR. Borman and Hewes (2002) found that Success for All students had better achievement outcomes, fewer special education placements and less frequent retention in grade, at a cost that was essentially the same as that of a control group. Borman and Hewes represented the cost-effectiveness of Success For All as effect size per $1000 of annual per pupil expenditures, finding that Success For All was more cost effective than three alternatives (STAR class size reduction, Perry Preschool, and Abecedarian Preschool) for which there are cost-benefit data.

Yeh (2007) compared the cost-effectiveness of a commercial Reading assessment program to four school improvement policies (a 10% increase in educational spending, voucher programs, charter schools and external accountability). Yeh represented cost-effectiveness as the ratio of effect size divided by cost, finding that the Reading assessment program was dramatically more cost-effective than the alternatives.

Creemers and van der Werf (2000) examined the cost-effectiveness of an Indonesian program that integrated teacher development, educational management, learning materials and community participation. They represented cost-effectiveness as the ratio of effect size divided by cost, defining a ratio of .0025 as substantial. They concluded that the program was expensive, requiring an increase in annual per pupil expenditures of 50% to increase student achievement by one standard deviation.

Of the CSR programs for which cost-effectiveness data are available, the closest approximation to Struggling Schools is Success For All, one of the most extensively implemented and investigated CSR approaches. Rigorous, evaluations of Success For All report statistically significant impacts on reading skills (Borman, Hewes, Overman, & Brown, 2003; Borman, Slavin, Cheung, Chamberlain, Madden, & Chambers, 2007). Success For All is a whole school intervention that focuses on the improvement of literacy (and to a lesser extent mathematics) in K-5. The program combines diagnostic assessment, cross-age grouping, Reading tutors, cooperative learning, and intensive instruction on discrete skills, supplemented by Family Support Teams to address out-of-school impediments to learning. Struggling Schools is similar to Success For All in that it was implemented in schools serving disadvantaged and underachieving student populations; instructional interventions focused on the development of reading skills using evidence-based practices; diagnostic assessment and differentiation of instruction were core features. Struggling Schools differed from Success For All in that it was much less prescriptive: Success For All provides curriculum materials, detailed instructional techniques, and tight implementation protocols, themes that were absent in Struggling Schools. In addition, Struggling Schools did not provide the social service component found in the Family Support Teams of Success For All. The fundamental difference between the two interventions is

that Success For All represents an externally developed program approach to CSR and Struggling Schools represents a CSR approach of internally developed school improvement process with support from outside agencies.

Although there were differences in the specific statistics of comparison, each of the studies of CSR cost-effectiveness reviewed above used a version of cost per standard deviation of improvement to determine cost-effectiveness. We selected the metric of effect size per $1000 to produce this decision rule: Struggling Schools would be cost-effective if (i) it increased the number of students reaching the provincial achievement standard, and (ii) did so at an effect size cost per $1000 no greater than Success For All.

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 for the full study 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? (3) Was the program cost-effective? The focus of this article is Question (3). The impact of Struggling Schools on achievement will be briefly presented to provide the outcome data used in the cost-effectiveness analysis. The detailed achievement analysis is in Authors (2009).

Methodology

The design of the study was a quasi-experimental, pre-post matched sample design with school as unit of analysis.

Sample

For each of the 100 schools participating in Struggling Schools we selected a control school that was not involved in the program. For each treatment school we identified all schools in the province that were within 0.5 SD of its prior Grade 3 Reading achievement. We selected as the matching control the school that was closest on composite SES (socio-economic status) score. Because the assessment agency suppresses achievement data from schools with less than 15 students in a grade, the sample reduced to 90 pairs. Compared to the 4054 elementary schools in the province, the treatment schools were at the 15th percentile in average income, at the 25th percentile in percentage of parents with some university, and at the 85th percentile in proportion of aboriginals, single parents, and unemployed.

Sources of Data

Achievement.

Grade 3 Reading scores came from criterion referenced assessments conducted by an agency independent of the provincial government. Responses to 32 multiple choice and 12 open ended items were aggregated to produce a 0-4 global score. The measure was the proportion of students in each school, with no exemptions, who achieved the provincial standard (level 3 or 4). The assessment agency used three procedures to ensure reliability: 1) group marking: during training, all markers scored the same student and discussed the results; 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. 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.

SES

We calculated a composite SES score for each of the 4,054 schools in the province using 14 variables in the census database and the formula of Johnson (2005). The SES composite was used to select the control group sample and was a covariate in the MANCOVA designed to reduce the disturbance of SES on the impact of the Struggling Schools program.

Costs of Struggling Schools

We used a combination of budgetary and opportunity costs to calculate the costs of Struggling Schools. The opportunity costs were generated using the ingredients method (Levin & McEwan, 2001). We estimated four types of costs. First, personnel and facilities costs were mainly for personnel assigned to the project full time (leadership advisors, case managers, diagnosticians, administrative assistants) or part time (program manager and portions of senior staff time). We used the mid-point of the salary range for each position X percentage time allocated to the program. We included the costs of benefits (22.3%) on salaries and the market costs for Ministry facilities (rooms for meetings held at Ministry offices). Costs for senior Ministry staff (manager, coordinators and education officers) were calculated at $155 per hour; costs for other Ministry personnel were calculated at $43.90 per hour. The cost of facilities was less than 1% of the total cost because most of the meetings were held at schools. Ministry conference rooms were calculated at $275 per day, the mid-point ($200-350) for conference room rentals in major hotels in the provincial capital. Following Levin and McEwan (2001) we did not include facilities costs for events hosted at schools because the cost of these rooms were included in the basic PPE (Per Pupil Expenditures) for all provincial schools. Consumables (office supplies, resource materials), technology requirements, and travel costs (mileage, airfares, hotel and meal allowances) were based on Ministry records and staff estimates. We estimated the costs of Steering Committee personnel based on per diem rates paid by the Ministry: $600 per

day for managers and $250 per day for non-managers. In addition to these operational costs there were expenses for monitoring the program, including formal evaluations contracted with outside agencies.

Second, grants to schools were measured using Ministry records of direct payments to schools for teacher release time, leadership training, professional learning materials and student resources. Although some educational economists are reluctant to use expenditure statements to calculate program costs, some high quality studies do so. For example, McEwan and Carnoy (2000) estimated the costs of a voucher system in Chile by examining sources of school revenue such as voucher payments and contributions to the schools by municipalities. Objections to including budgetary allocations when estimating costs include: (i) costs for a specific program are embedded in the costs of a larger unit of operation and are difficult to disentangle (Levin & McEwan, 2001); (ii) expenditures may underestimate costs if the program shifts costs from paid staff to unpaid volunteers (such as parents, as in King, 1994); (iii) expenditures may over-estimate costs if a capital cost is assigned to a single year when the facility will be used for many years (Harris, 2009). None of these conditions applied in our study: (i) grants to schools were clearly distinguished from other Ministry operations and schools were visited at least monthly to ensure that the funds were expended for program purposes. (ii) The opportunity costs of donated time were calculated and included in the costs of the program. (iii) Capital costs (including professional learning materials for teachers that could be used beyond the life of the project) were negligible.

Third, we included estimates of unfunded school costs; i.e., the market value of donated time. Program implementation created additional uncompensated workload for teachers (eight hours per week in the first two years and two hours per week in the third year) and principals (24

hours per month each year), estimated at $200 per day for teachers and $250 for principals. Costs for teachers were calculated using supply teacher (replacement) rates. Benefits for teachers and principals were not added because these costs were included in the PPE for elementary schools. Student time was not included as a cost. School costs were obtained through interviews with teachers and principals in case study schools. We converted all costs to 2006 constant dollars using the Bank of Canada inflation calculator; the exchange rate in 2006 was CAN$1.00=US$0.85.

Fourth, we added these incremental costs of Struggling Schools to the average per pupil expenditures (PPE) for elementary schools in the province, which was the sum of a foundation grant, special purposes grants, and pupil accommodation grants. The Ministry provided the PPE for 2005-06 and we inflated the PPE for 2006-07 by 2.41%.

Control group school costs were the average PPE for elementary schools. At the time of the study there was very little between-school variation in PPE. There was a learning opportunities grant to enable school districts to provide programs for low achieving students. The grant was based on socio-economic factors from Statistics Canada (the exact formula is not published but it is based on predictors of academic difficulty: proportion of residents in the district who are low income, low education, recent immigrants, and aboriginals). The size of the grant varied from 0.47% to 5.5% of a district's total budget. Since the treatment and comparison schools were not significantly different on SES profile, the PPEs for their districts would have been increased by the learning opportunities grant by similar percentages. But the grant is awarded to the district, not the school. And although districts are audited for compliance with provincial regulations, they are not required to direct the funds from the learning opportunities grant to the schools with the greatest proportion of at-risk pupils.

Costs of Success For All

Cost data for Success For All were drawn from the analysis of costs in five Baltimore Elementary schools provided by Borman and Hewes (2002). Four components were summed: (i) The ingredients model was used to estimate marginal costs. The Success For All Foundation identified the ingredients in the program and provided standard estimates of salary and benefits for reading tutors, family support staff, and in-school program facilitators, to which were added training, implementation, and curriculum materials costs. These costs were estimated for each school (three of the five schools did not hire Family Support staff) and aggregated across schools and eight years of the program. (ii) Per Pupil Expenditures based on the annual current expenditures per pupil for all American schools for the 1999-2000 school year. (iii) An estimate of the market value of special education services was based on 1985-86 data for non-severe handicaps. (iv) The cost of retention in grade was based on the average PPE for Success For All schools and discounted by 5% for control group schools.

In calculating effect size per $1000 of program cost, Borman and Hewes included only (i) the marginal costs, adjusted by (iii) the cost of special education services and (iv) the cost of retention in grade. The annual PPE of Success For All was multiplied by the average number of years students were in the program. This procedure omitted (ii) the annual current PPE for all schools: cost-effectiveness was defined as effect size per $1000 of incremental cost. In comparing Struggling Schools to Success For All we followed the same procedure except that we did not include the cost of special education services or the cost of grade retention. We omitted these because they are not an integral part of Success For All (see Borman & Hewes, 2002, footnote 8) and represented only 1.7% of the total program cost. In addition, for Struggling

Schools special education costs were embedded within the annual PPE grants to schools and retention in grade was virtually zero because it was and is strongly discouraged by the province.

Analysis

After demonstrating the equivalence of the Struggling Schools and control samples on all measured variables, we conducted a multivariate analysis of covariance using GLM (General Linear Modeling). The dependent variables were grade 3 Reading scores in 2005-06 and 2006-07. The covariates were prior achievement in Grade 3 Reading and school SES. The independent variables were group (treatment or control), program Phase (years in the program), and the group X Phase interaction. This design had adequate statistical power for determining whether there was an overall effect of the Struggling Schools program. With a sample size of 180, 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.

Our source for the effectiveness of Success For All was the quasi-experiment reported by Borman and Hewes (2002). Volunteer schools were matched on demographic variables with control schools. Student outcomes were scores on grade 8, standardized measures of Reading skills, adjusted by SES and Kindergarten achievement, using multi-level modeling.

To determine cost-effectiveness we calculated the benefits and costs of each phase of the Struggling Schools program and its control group. For each group of schools we calculated the average increase in PPE required to bring one student to the provincial achievement standard, i.e., we divided the annual PPE by the adjusted mean achievement level.

Our first decision rule was that Struggling Schools was cost-effective if it increased the number of students reaching the provincial achievement standard and did so at a cost per successful student that was no more than 25% greater than the cost per successful student in the

control. We conducted sensitivity analysis by varying our estimates of costs (by deleting donated time) and benefits (by raising the effect of the program from the mean to the upper bound of the 95% confidence level). We also calculated the cost-effectiveness of the Struggling Schools program in comparison to Success For All by converting the marginal cost of Struggling Schools over four years of funding to 2000 US$ (1 US$=0.67 $CAN in 2000; 2000-2006 inflation rate=11.4%) and dividing the total incremental cost of Struggling Schools by its effect size. This enabled us to compare the effect size per $1000 of Struggling Schools to the effect size per $1000 for Success For All reported in Borman and Hewes (2002).

Results

Achievement Effects of Struggling Schools

Table 1 shows the effect sizes, with bias correction for small samples (Hedges & Olkin, 1985), and the lower and upper bounds of the effect sizes (95% confidence level). The table also shows how long each cohort was in the program at the time of the data collection. There was a statistically significant program effect when all Phases were aggregated; there was no significant Phase or Phase X Treatment effect. Details of the MANCOVA and other procedures are reported in Authors (2009). The effect size of the program improved slightly, from *d*=.48 in 2005-06 to *d*=.60 in 2006-07. But the larger effect size was the result of a smaller pooled standard deviation in 2006-07 (.13) than in 2005-06 (.17). Overall there was no improvement in the mean achievement score of either treatment or control group schools from 2005-06 to 2006-07. Table 1 suggests the effect of the program was highest in the third program year, declined through the exit year and was still positive two years from exit.

Table 1 about here

Cost-effectiveness of Struggling Schools

Table 2 summarizes the costs of the program in constant 2006 dollars for the schools in each Phase of the program, by academic year. Grants to schools are funds transferred to implement school improvement plans. For schools in Phases 1, 2 and 3 the grants were approximately $300,000 (in nominal dollars), varying by school size, across the three years of the program. For schools in Phase 4, the grants were reduced to approximately $200,000 (in nominal dollars) and the distribution of funds was 52%, 21% and 21% across the three years, rather than the one-third distribution per year for Phases 1-3. Schools in their fourth (exit) year received no grants and no teacher time was charged; personnel and facilities costs continued because the schools continued to interact with program staff. Cumulative total costs for 2000-2006 and 2000-2007 are shown by Phase in the last two rows.

Table 2 about here

Table 3 summarizes the annual PPE per successful student in program and control schools, calculated by dividing the PPE for the group (treatment or control) by its achievement mean. For example, for Phase 1 Treatment schools, the total cost over four years of the program, $10,367,812 (from Table 2) was divided by the number of pupils receiving the program (1.5 classes x 30 students x 15 schools) and by the number of years in the program (4); these total marginal costs of the program, $10,367,812/(1.5 * 30 *1.5 * 4)=$3839.93, were added to the average cost of elementary schools in the province not receiving the program; i.e., $3839.93 + $8193=$12,032.93. The achievement means were adjusted by prior school achievement and composite school SES. In six of the eight comparisons, the annual PPE per successful student was more than 25% higher for schools in the Struggling Schools program than for control schools. Averaging across phases, the PPE per successful student was 36% higher in 2005-06

and 26% higher in 2006-07 than in the control schools. Struggling Schools did not meet our first criterion for cost-effectiveness.

Table 3 about here

In Table 4 we report the results of a sensitivity analysis in which we varied our assumptions about costs and benefits. First, we created lower cost estimates by removing donated time (i.e., additional teacher and principal time required by program implementation). The fourth column of Table 4 show that in all but one of the eight comparisons the annual PPE per successful student of Struggling Schools was less than 25% more than in the control schools. The sixth column of Table 4 shows the results when we used the upper bound of the adjusted achievement means from the MANCOVA and the lower estimate of program costs. In all of the eight comparisons the Struggling Schools program was less than 25% more costly than the status quo. In two of the comparisons for 2006-07, the cost per successful student was lower in Struggling Schools than in control schools. The sensitivity analysis suggests that we should be cautious in our initial claim that Struggling Schools was not cost-effective. Changing assumptions behind the calculation of costs and benefits supports the claim that Struggling Schools was cost-effective.

Table 4 about here

Finally, in Table 5 we summed the total increase in per pupil expenditures across the four years that schools were in the Struggling Schools program, converting the totals to 2000 US$. We extrapolated the exit year costs for Phase 4 schools from Phase 3 schools data. We calculated the 2005-06 and 2006-07 effect size per $1000 of Struggling Schools funding and compared this effectiveness-cost ratio to the same ratio for Success For All schools, based on students being in Success For All for an average of 3.84 years. The overall ratio for Struggling Schools was 0.06

in 2005-06 and 0.07 in 2006-07, compared to 0.09 for Success For All. In terms of our second criterion, Struggling Schools program was not cost-effective.

Table 5 About Here

Discussion

We developed two standards for determining whether Struggling Schools was cost-effective. First, we compared the program's cost-effectiveness to that of control schools: the Struggling Schools program would be cost-effective if (i) it increased the number of students reaching the provincial achievement standard, and (ii) did so at a cost per successful student of no more than 125% of the cost per successful student of similar students in control schools. The program met part (i) but not part (ii). Second, we compared the program's cost-effectiveness to Success For All, an externally developed CSR programs with similar objectives: The Struggling Schools program would be cost-effective if (i) it increased the number of students reaching the provincial achievement standard, and (ii) did so at a ratio of effect size cost per $1000 of increased PPE that was no greater than of Success For All. Again, the program met part (i) but not part (ii). We conclude, reluctantly and cautiously, that Struggling Schools was not cost-effective.

Comparisons of Struggling Schools to Control Schools

In our sensitivity analysis we found that Struggling Schools would have reached our first cost-effectiveness criterion if donated time of teachers and administrators had been excluded from cost calculations. The issue is controversial. The costs of school improvement are typically borne by school staff, are frequently unrecognized by policy makers, and are difficult to measure. Slavin and Madden (2003) argued that "most Success For All schools never have received funds beyond their usual Title I allocations, so in one sense the program has no

incremental costs" (p. 4), an argument that could be applied to the donated time required to implement Struggling Schools. Among economists, Levin (2002) argued that when assessing costs in CSR programs donated ingredients should be included because it does not matter who provides the resources: they are still costs. Harris (2009) concurred, suggesting that the distribution of costs among stakeholders is unimportant because these will be negotiated through political bargaining. In contrast, King (1994) argued that donated time could be considered cost-free if it is readily available. She suggested that in a community that provides high levels of donated personnel time, Success For All is inefficient because it replaces donated time with paid staff. The cost of Struggling Schools, with regarded to donated time, is not fixed: this component may be a large or small addition to the incremental cost of the program, depending upon the culture of the school. In addition, the willingness of staff to donate the time is likely to be a function of their commitment to the school improvement process. Since donated time made the difference between meeting or not meeting our first decision rule, investigation of the typicality of the donated time estimates provided by teachers and students in our case studies is warranted.

Our first decision rule recognized that to raise the number of students meeting proficiency standards becomes increasingly challenging as one moves to lower levels of the student ability pool. The relationship between cost and success is non-linear; i.e., there is a threshold of minimal resources required for even the most able learners, a steady increase in resources to meet students of increasing challenge, and ceiling effects when further resources fail to contribute to more students reaching proficiency. We drew upon mastery learning research to suggest that raising the performance of the harder-to-educate group should not be more than 25% of the costs of success in control schools. But the analogy to mastery learning is imperfect: the

assessments were aligned to the province's curriculum objectives which emphasized Reading comprehension as well as mastery of discrete skills.

Comparison of Struggling Schools to Success For All

Our second decision rule involved a comparison to one of the most extensively investigated CSR programs, Success For All. Struggling Schools cost estimates included costs of program development, e.g., costs incurred before schools were identified. In contrast, the cost-effectiveness calculations for Success For All were based solely on delivery costs after the program had been fully developed and extensively field tested. During the roll out of Struggling Schools, a major portion of its costs (grants to schools) were reduced substantially for the last cohort of schools (Phase 4) with no loss of achievement benefits and the costs of program management were distributed across more schools than in Phases 1-3. Struggling Schools was becoming more cost-effective as it was scaled up. But in the user-generated approach to CSR, the processes of school improvement have to be enacted anew in each school. What is exported to new sites is not the products of innovation but the innovation process (Fullan, 1999). CSR approaches in which schools select from a menu of previously tested options are inherently more cost-effective than user-generated programs developed from school improvement principles.

Cost-effectiveness is dependent in part on how outcomes are calculated. In Struggling Schools student outcomes were aggregated to the school level by the provincial assessment agency whereas in Success For All achievement was reported at the individual student level. Since differences between-schools are usually smaller than differences within-schools, the pooled standard deviation will be smaller at the school level and the effect sizes will be larger (Hall, Tickle-Degnen, Rosenthal, & Mosteller, 1994). The cost-effectiveness advantage of Success For All over Struggling Schools was likely greater than we detected.

The impact of programs varies from one site to another. Borman and Hewes (2002) found an effect size *d*=.29 for Success For All in five Baltimore schools, which was 60% higher than the *d*=.18 for the 42 Success For All studies in Borman et al. (2003) and 260% higher than the *d*=.08 reported in the same meta-analysis for Success For All studies conducted by third-party evaluators. In contrast, Slavin and Madden (2000) reported *d*=.39-.62 for studies of Success For All conducted from 1988 to 1999. Since Success for All is highly standardized, cost data could be estimated for these outcome studies, producing a less precise but reasonable accurate cost estimate. But the cost-effectiveness ratio would vary in response to fluctuations in effect sizes. Whether Struggling Schools was deemed to be cost-effective could depend upon which Success For All study it was compared to.

This problem is not limited to comparisons to Success For All. For example, Tennessee STAR, widely recognized as "one of the great experiments in education in U.S. history" (Mosteller, Light, & Sachs, 1996, p. 814), provides the data on which most claims about the cost-effectiveness of class size reductions are based (Borman et al., 2005; Borman et al., 2007; Harris, 2009; Krueger, 2003; Levin, 2009; Yeh, 2007). But the effects of class size on achievement are highly variable. Much lower estimates of the effects of class size reduction were reported by Stecher, Bohrnstedt, Kirst, McRobbie, and Williams (2001) for California and Hruz (2000) for Wyoming. It may be that so few evaluations of CSR programs contain sufficient cost data to make cost-effectiveness comparisons that the bar is set very high.

Only student achievement was included in determining the cost-effectiveness of Struggling Schools and the CSR program to which it was compared. There was no consideration of teacher capacity impacts (such as increased use of evidence-based instructional skills and improvements in teacher efficacy) or organizational capacity effects (such as the strengthening of

professional communities and movement toward distributed or transformational leadership strategies associated with school improvement as found by Authors, 2006; Ylimaki, Jacobson, & Drysdale, 2007). Although the ultimate criterion of school success is student achievement, the user generated approach to CSR assumes that achievement will improve in the long run if the meditational effects of teacher and organizational capacity enhancements are included in the assessment model.

Directions for Future Research

In 2006-07 the Struggling Schools program was renamed, restructured to lower costs, and scaled up to 800 schools. In 2007-08 and in 2008-09, the principles embedded in the Struggling Schools program were incorporated into a school improvement process rolled out to all schools in the province. These successor programs are much less costly than Struggling Schools but their benefits are unknown. Research is needed to determine whether the effect sizes for the original Struggling Schools program are sustained in the scaled up versions.

More generally there is a need for more research on the cost of CSR programs. Most CSR researchers are focused solely on program benefits, i.e., whether the program generated a statistically significant effect of meaningful size. But policy makers and administrators constantly ask the value for money question. When considering alternative ways of addressing school needs, they make cost-effectiveness judgments with little to go on beyond their intuitions. Cost-effectiveness studies need to be completed and reported in evidence banks such as http://ies.ed.gov/ncee/wwc/ and http://www.bestevidence.org.

Conclusion

Our study of the cost-effectiveness of Struggling Schools makes several contributions to CSR research. First, it is a fresh case. There are very few studies of the cost-effectiveness of CSR

and virtually all draw upon secondary sources rather than collecting new data. In addition, the Struggling Schools program is typical of CSR programs developed by states and districts to increase the capacity of low achieving schools serving disadvantaged populations. The principles embedded in Struggling Schools' program theory and the structures and processes created to build school capacity are similar to those put in place in countless jurisdictions. The study found that the approach to CSR represented by Struggling Schools made a statistically significant contribution to student achievement of moderate effect size, confirming the worth of such programs. The distinctive finding of our study is that the cost of doing so was high. The program was not cost-effective when compared to the status quo or to a frequently implemented alternative CSR program, Success For All. But our claim is offered with serious caution: different choices about costs and the program to which it is compared could have led to a claim that Struggling Schools was cost-effective. The practical implication is that states and districts developing similar programs need to be very focused on costs, particularly when scaling up the innovation across a large number of schools.

The second contribution of the study is the finding that when comparing Struggling Schools to other CSR programs, selecting from a menu of CSR options is likely to be more cost-effective than developing a new program. This finding is likely to generalize because schools that select from a menu of options do not bear development costs.

The third contribution of the study is that the value for money question raises additional unresolved questions that warrant the attention of school improvement researchers. Should unfunded teacher and principal costs of implementation be included or should school personnel be expected to eat the costs because schools are expected to be continuously improving? When calculating the benefits of CSR, should researchers focus on achievement alone or include

teacher and organizational learning outcomes as well? If so, how should these very different benefits be weighed if they cannot be transformed into monetary values?[1] To which programs should the cost-effectiveness of particular CSR programs be compared—the status quo, the best, or the typical? We cannot begin to answer these questions unless researchers make the investigation of CSR costs as important as the study of benefits.

Endnotes

1. There are procedures for combining multiple program outcomes into a single cost-effectiveness analysis. Levin & McEwan (2001) suggest several strategies based on multi-attribute utility analysis. These procedures require substantial additional data, i.e., evidence of the effects of the program on other outcomes (such as impacts on instructional practices, professional learning communities and leadership styles) and information on the value preferences of stakeholders, quite apart from the challenge of assigning costs within the treatment and control to these specific benefits.

References

Arlin, M. (1984). Time variability in mastery learning . *American Educational Research Journal, 21*(1), 103-120.

Arlin, M., & Webster, J. (1983). Time costs of mastery learning. *Journal of Educational Psychology, 75*(2), 187-195.

Borman, G. D., & Hewes, G. M. (2002). The long-term effects and cost-effectiveness of Success For All. *Educational Evaluation and Policy Analysis, 24*(2), 243-266.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73*(2), 125-230.

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success For All. *American Educational Research Journal, 44*(3), 701-731.

Boyle, P. J., DuBose, E. R., Ellingson, S. J., Guinn, D. E., & McCurdy, D. B. (2001). *Organizational ethics in health care: Principles, cases, and practical solutions*. San Francisco: Jossey-Bass.

Creemers, B., & van der Werf, G. (2000). Economic viewpoints in educational effectiveness: Cost-effectiveness analysis of an educational improvement project. *School Effectiveness and School Improvement, 11*(3), 361-384.

Fullan, M. G. (1999). *Change forces: The sequel*. London: Falmer.

Fullan, M. G. (2002). *Leading in a culture of change*. San Francisco: Jossey-Bass.

Fullan, M. G. (2005). Turnaround leadership. *Educational Forum, 69*(2), 174-181.

Fullan, M. G., & Campbell, C. (2007, April). *The Ontario literacy and numeracy strategy.* Paper

presented at the annual meeting of the American Educational Research Association, Chicago.

Gettinger, M. (1985). Time allocated and time spent relative to time needed for learning as determinants of achievement. *Journal of Educational Psychology*, *77*(1), 3-11.

Hall, J. A., Tickle-Degnen, L., Rosenthal, R., & Mosteller, F. (1994). Hypotheses and problems in research synthesis. In H. Cooper, & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 17-28). New York: Russell Sage Foundation.

Harris, D. N. (2009). Toward policy-relevant benchmarks for interpreting effect sizes: Combining effects with costs. *Educational Evaluation and Policy Analysis*, *31*(1), 3-29.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hruz, T. (2000). *The costs and benefits of smaller classes in Wisconsin: A further evaluation of the SAGE program.* Thiensville, WI: Wisconsin Policy Research Institute Report.

Hummel-Rossi, B., & Ashdown, J. (2002). The state of cost-benefit and cost-effectiveness analyses in education. *Review of Educational Research*, *72*(1), 1-30.

Johnson, D. R. (2005). *Signposts of success: Interpreting Ontario's elementary school test scores*. Toronto: C.D.Howe Institute.

King, J. A. (1994). Meeting the educational needs of at-risk students: A cost analysis of three models . *Educational Evaluation and Policy Analysis*, *16*(1), 1-19.

King Rice, J. (1997). Cost analysis in education. Paradox and possibility. *Educational Evaluation and Policy Analysis*, *19*(4), 309-318.

Krueger, A. B. (2003). Economic considerations and class size. *Economic Journal*, *113*(485), F34-F63.

Kulik, C.-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60*(2), 265-299.

Levin, H. M. (2002). *The cost effectiveness of whole school reforms*. Retrieved May 4, 2009, from http://www.cbcse.org/media/download_gallery/Whole%20School%20Reforms.pdf.

Levin, H. M. (2009). The economic payoff to investing in educational justice. *Educational Researcher, 38*(1), 5-20.

Levin, H. M. (2001). Waiting for Godot: Cost-effectiveness analysis in education. *New Direction for Program Evaluation, 90*, 55-68.

Levin, H. M., Glass, & Meister, G. (1987). A cost-effectiveness analysis of computer-assisted instruction. *Evaluation Review, 11*, 50-72.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications (2nd edition)*. Thousand Oaks, CA.: Sage.

Martinez, J. G. R., & Martinez, N. C. (1999). Teacher effectiveness and learning for mastery. *Journal of Educational Research*; 92, 279-285.

McEwan, P. J., & Carnoy, M. (2000). The effectiveness and efficiency of private schools in Chile's voucher system. *Educational Evaluation and Policy Analysis, 22*(3), 213-239.

Mosteller, F., Light, R. J., & Sachs, J. (1996). Sustained inquiry in education: Lessons learned from skill grouping and class size. *Harvard Educational Review, 66*, 797-842.

Ross, S. M., Alberg, M., & Nunnery, J. (1998). *Selection and evaluation of locally developed versus externally developed schoolwide projects.* Memphis, TN: University of Memphis.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime effects:* The High/Scope Perry Preschool Study Through Age 40 (Monographs of the High/Scope Educational Research Foundation). Ypsilanti, MI : High Scope Press.

Slavin, R. E., & Madden, N. A. (2000). Research on achievement outcomes of Success for All: A summary and response to critics. *Phi Delta Kappan, 82*(1), 38-40, 66-59.

Slavin, R. E., & Madden, N. A. (2003). *Scaling up Success For All: Lessons for policy and practice*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Stecher, B., Bohrnstedt, G., Kirst, M., McRobbie, J., & Williams, T. (2001). Class-size reduction in California: A story of hope, promise, and unintended consequences. *Phi Delta Kappan, 82*(9), 670-674.

Yeh, S. S. (2007). The cost-effectiveness of five policies for improving student achievement. *American Journal of Evaluation, 28*(4), 416-436.

Ylimaki, R. M., Jacobson, S. L., & Drysdale, L. (2007). Making a difference in challenging, high-poverty schools: Successful principals in the USA, England, and Australia. *School Effectiveness And School Improvement, 18*(4), 361-381.

Table 1

*2005-06 and 2006-07 Unadjusted Grade 3 Reading Achievement Means and Standard Deviations, by Group and Year in Program**

	Treatment Schools			Control Schools			ES Confidence Interval		
	N	Mean	SD	N	Mean	SD	ES	Lower	Upper
2005-06 Reading									
Phase 1 (exit year + 1)	14	.50	.17	14	.45	.18	.28	-.47	1.02
Phase 2 (exit year)	13	.48	.16	13	.44	.15	.25	-.52	1.02
Phase 3 (year 3)	11	.62	.22	11	.48	.11	.77	-.09	1.64
Phase 4 (year 2)	52	.54	.17	52	.47	.16	.42	.03	.81
All Phases	90	.54	.18	90	.46	.15	.48	.18	.78
2006-07 Reading									
Phase 1 (exit year + 2)	14	.50	.16	14	.47	.17	.18	-.57	.92
Phase 2 (exit year + 1)	13	.47	.21	13	.43	.12	.23	-.54	1.00
Phase 3 (exit year)	11	.56	.18	11	.49	.20	.35	-.49	1.20
Phase 4 (year 3)	52	.56	.17	52	.45	.16	.66	.27	1.06
All Phases	90	.54	.10	90	.46	.16	.60	.30	.90

*From Authors (2009)

Table 2

Summary of Program Ingredients and Costs in 2006 Dollars, by Year and Phase

	Phase 1	Phase 2	Phase 3	Phase 4
		Personnel & Facilities		
2000–01	$32,618	$30,443	$30,443	$123,947
2001–02	$439,055	$0	$0	$0
2002–03	$543,328	$507,596	$0	$0
2003–04	$1,133,657	$1,057,322	$1,057,322	$0
2004 -05	$503,324	$469,769	$469,769	$1,912,632
2005–06	$0	$469,769	$469,769	$1,912,632
2006–07	$0	$0	$469,769	$1,912,632
		Unfunded School Costs		
2000–01	$0	$0	$0	$0
2001–02	$1,065,787	$0	$0	$0
2002–03	$1,049,244	$979,294	$0	$0
2003–04	$346,837	$956,088	$956,088	$0
2004 -05	$340,062	$317,391	$937,411	$3,816,603
2005–06	$0	$313,040	$313,040	$3,764,280
2006–07	$0	$0	$313,040	$1,274,520
		Grants to Schools		
2000–01	$0	$0	$0	$0
2001–02	$2,882,982	$0	$0	$0
2002–03	$1,027,636	$2,566,407	$0	$0
2003–04	$1,003,284	$907,197	$2,421,229	$0
2004 -05	$0	$889,476	$859,527	$6,613,743
2005–06	$0	$0	$847,744	$2,361,802
2006–07	$0	$0	$0	$2,361,802
		Cumulative Total Cost		
2000-06	$10,367,812	$9,463,791	$8,362,342	$20,505,639
2000-07	$10,367,812	$9,463,791	$9,145,151	$26,054,593

Table 3

Cost of Reaching the Provincial Achievement Standard in Treatment and Control Schools by Phase, for 2005-06 and 2006-07

	Control Schools			Treatment Schools		
	Annual PPE	Adjusted Achievement*	Annual PPE per Successful Student	Annual PPE	Adjusted Achievement*	Annual PPE per Successful Student
	2005-06 Results (in 2006 dollars)					
Phase 1	$8,193	.45	$18,207	$12,033	.52	$23,140
Phase 2	$8,193	.48	$17,069	$12,887	.52	$24,783
Phase 3	$8,193	.49	$16,720	$13,724	.62	$22,135
Phase 4	$8,193	.45	$18,207	$13,765	.54	$25,491
	2006-07 Results (in 2006 dollars)					
Phase 1	$8,390	.46	$18,239	$11,590	.56	$20,696
Phase 2	$8,390	.46	$18,239	$12,145	.51	$23,814
Phase 3	$8,390	.52	$16,135	$12,926	.57	$22,677
Phase 4	$8,390	.43	$19,512	$12,622	.55	$22,949

PPE=per pupil expenditures over the three years of the program
* Mean achievement scores were adjusted by school prior achievement and school SES composite. The covariates were evaluated at prior achievement=.2787 and SES=10.0827 in the formula: post achievement= intercept + group * phase + group + phase + prior achievement + SES.

Table 4

PPE per Successful Student in Program and Control, Assuming Lower Cost and Assuming Lower Cost and Higher Benefit, by Phase

	Annual PPE per Successful Student				
		Lower Cost		Lower Cost/Higher Benefit	
Program Phase	Control	PPE	% Above	PPE	% Above
	2005-06 Results				
Phase 1 (exit year + 1)	$18,207	$21,144	16%	$18,325	<1%
Phase 2 (exit year)	$17,069	$22,336	31%	$19,040	12%
Phase 3 (year 3)	$16,720	$19,781	18%	$17,033	2%
Phase 4 (year 2)	$18,207	$22,070	21%	$20,548	13%
	2006-07 Results				
Phase 1 (exit year + 2)	$18,239	$19,152	5%	$16,758	-8%
Phase 2 (exit year + 1)	$18,239	$21,818	20%	$18,859	3%
Phase 3 (exit year)	$16,135	$20,484	3%	$17,691	10%
Phase 4 (year 3)	$19,512	$20,335	10%	$18,955	-3%

PPE=per pupil expenditures over the three years of the program

Table 5

Marginal Per Pupil Expenditures of Struggling Schools and Success For All, in 2000 US$

CSR Program	Marginal PPE	Reading ES	Effect per $1000
	Struggling Schools 2005-06		
Phase 1	$11,379	0.28	0.03
Phase 2	$11,129	0.25	0.02
Phase 3	$9,834	0.77	0.08
Phase 4	$5,923	0.42	0.07
All Phases	$8,018	0.48	0.06
	Struggling Schools 2006-07		
Phase 1	$11,379	0.18	0.02
Phase 2	$11,129	0.23	0.02
Phase 3	$10,754	0.35	0.03
Phase 4	$7,525	0.66	0.09
All Phases	$9,060	0.60	0.07
	Other CSR Program		
Success For All	$3,054	0.29	0.09

PPE=per pupil expenditures over the four years of the program for Struggling Schools and over 3.84 years for Success For All. Success For All data from Borman & Hewes (2002).