An Evaluation of the "100 Book Challenge Program"

October, 1999

Dr. Joseph DuCette
Department of Educational Psychology
Temple University

Executive Summary

This report presents a series of analyses on data collected to evaluate the "100 Book Challenge Program". The "100 Book Challenge Program" is designed to help students in the School District of Philadelphia improve their level of proficiency in reading; An integral part of the Program is the stipulation that students read a minimum of 100 books during a designated period of time. Initially, the Program was implemented for approximately 14 months in the first, second and third grades in twelve elementary schools in the District.

The evaluation utilized a quasi-experimental design in which students in Program classes were compared to students at the same grade and the same school who were not attending Program classes. The Stanford Achievement Test 9th Edition (SAT-9) was used to assess the reading level of Program and comparison students. The SAT-9 was given at the end of the 1998–1999 academic year.

The results of the evaluation indicated that:

- Program students attained significantly higher levels of reading achievement as compared to students in non-Program classes. This was true for students in all three of the grades tested, with the largest differences occurring for students in the first and third grades.
- Program students in classes with a higher degree of Program implementation attain significantly higher reading scores than students in classes with a lower degree of implementation.
- First grade students in Program classes are reading at higher levels (using the District's classification of "Below Basic", "Basic", "Proficient" and "Advanced") as compared to non-Program students.

Overall, the evaluation indicates clear support for the "100 Book Challenge Program". Suggestions for future research on the Program are provided at the end of the Report.

An Evaluation of the "100 Book Challenge Program"

Dr. Joseph DuCette
October, 1999

A · Introduction

Presented in this report are data concerning the effectiveness of the "100 Book Challenge Program". Briefly, the "100 Book Challenge Program" (hereafter referred to as "the Program") is a motivational and instructional support system where students are asked to read a minimum of 100 books at levels appropriate to their age and reading level throughout a specific period of time. Students may read more than this, and data are collected through the Program on the number of children who meet the goal of 100 books read, as well as 200 and 300 books. Both students and parents are asked to verify that the books have actually been read. Teachers who participate in the Program are provided with an array of books at various difficulty levels for their classroom. The central purpose of the Program is to improve the reading achievement of participating students.

The "100 Book Challenge Program" has been implemented in the School District of Philadelphia for approximately fourteen months. Data have been collected on various aspects of the Program during this time. Most of this data collection has focused on process evaluation and implementation issues. Although informal evaluation data have been collected, a systematic outcome evaluation has not been conducted since the Program was still in its formative phase. With support from various sources, including the Eagles Youth Partnership and the William Penn Foundation, it was felt that an initial outcome evaluation might prove valuable to ascertain if the Program is attaining its goal of improving the reading achievement of participating students. The results of this evaluation are presented in this Report.

B · Methodology and Instrumentation

Data were collected from students in 12 elementary schools in the School District of Philadelphia which were participating in the Program. The general rule for participation is that at least 10 teachers in each school have agreed to participate in the Program. For each school, classrooms at the same grade level not participating in the Program were selected as the control or comparison group. The evaluation, therefore, employs what Campbell and Stanley (1966) call a quasi-experimental design. This design is both common and acceptable in the educational evaluation literature.

The data were collected at the end of the 1998–1999 academic year and consist of scores from the reading achievement section of the Stanford Achievement Test 9th Edition.

The Stanford-9 (or SAT-9 as the test is often known) is given to all District students from the 2nd grade on (although scores for the second grade are not reported as part of District statistics.) The data for this evaluation are derived from first, second, and third grade classes since these classes constitute the major focus of the Program at this time. Since first grade students do not participate in the Stanford-9 test administration program, a special administration of the open-ended part of the test was given to first grade students in the Program as well as a matched group of students not in the Program. There were some Program classes which were either non-graded, special support or where students' first language was not English. Since the Stanford-9 was not given in these classes, they will not be included in the analyses for the evaluation. Some data from these classes, however, will be presented for descriptive purposes. The core question being asked in this evaluation is whether the Program has had a significant impact on the reading achievement of students. Various analyses are presented in this paper to address this question.

The data from the Stanford-9 are expressed as Normal Curve Equivalent Scores (NCE's). NCE's are a form of derived scores which are standardized so that the average is set at 50. NCE's are similar to, but are not identical with, percentile rank scores. NCE's are most commonly used to report aggregated statistics on school districts or schools within districts. The Stanford-9 produces scores in three areas for reading: a multiple choice reading comprehension score; an open-ended reading comprehension score; and a composite reading score. All three of these scores were available for analysis for the third grade. Only the open-ended score was available for students in the first and second grades. Since the open-ended

score is the only piece of data available on all students, the major focus of this analysis will be on these data. In addition, since data were provided by classroom and not by individual student, the unit of analysis for all of the statistical procedures that will be presented in this report is the classroom. This is a recommended procedure for educational research of this type, although it must be emphasized that this method of analyzing data is conservative. That is, the sample size for the statistical procedures that will be presented in this report where the classroom is the unit of analysis is much smaller than would be the case if the individual student was the unit of analysis. This reduces power, but makes significant results more meaningful.

In addition to the SAT-9 reading achievement scores for students in Program and comparison classes, data were available for Program classes on the proportion of students who had attained the 100 book criterion, as well as the proportion of the students who had attained the criteria of 200 and 300 books. Moreover, an Implementation Score derived from direct classroom observation was available on most of the Program classes. This score represents the extent to which the Program was being fully implemented in each specific classroom. These data will be used for a within-Program evaluation.

Finally, as part of the special administration of the open-ended SAT-9 test for first grade students, the Program and comparison students were placed into reading levels

using the District's classification of "Below Basic", "Basic", "Proficient" and "Advanced". The two groups will be compared on their performance using this four-way classification system.

C · Sample Description

Table 1 presents the breakdown of classes and students for the Program and comparison groups by school. The data for the number of students includes only those students for whom data were available for analysis.

Table 1: Description of the Program and Comparison Groups by School

	Program		Comparison	
	Number of Classes	Number of Students	Number of Classes	Number of Students
Alcorn	10	165	10	203
Anderson	10	181	4	14
Drew	11	195	2	47
Emlen	10	191	3	47
Harrington	9	223	10	103
Hartranft	9	140	14	134
Lowell	10	206	5	132
Meade	10	204	9	161
Richmond	10	207	12	193
Southwark	12	133	2	5
Taylor	13	172	7	90
Whittier	13	144	1	27
Total	**127**	**2161**	**79**	**1156**

A description of the sample by group and grade level is presented in Table 2. This table includes only those grades for which evaluation data are available.

Table 2: Description of the Program and Comparison Groups by Grade Level

	Program		Comparison	
	Number of Classes	Number of Students	Number of Classes	Number of Students
Grade 1	38	839	9	148
Grade 2	37	817	31	505
Grade 3	22	464	24	357

D · Analysis of the Open-Ended Reading Comprehension Data

The means and standard deviations for the open-ended reading comprehension scores from the Stanford-9 by group are presented in Table 3. The Table also presents the results of a separate-sample t-test comparing the two groups.

Table 3: Means, Standard Deviations and t-test Results for AH Program and Comparison Students Across Grade Levels

	Mean	Standard Deviation	t-test	Significance of t-test
Program Classes	42.70	11.07	3.35	.001
Comparison Classes	35.74	14.33	3.35	.001

As evident from Table 3, the Program students have achieved a significantly higher NCE score on the open-ended section of the SAT-9 as compared to the non-Program students. To demonstrate this is a somewhat more complex manner, the open-ended scores were analyzed by group and grade level. These means are presented in Table 4.

Table 4: Means for the Open-Ended SAT-9 Reading Comprehension Test by Group and Grade

	Program Students	Comparison Students
Grade 1	43.72	33.16
Grade 2	43.22	38.49
Grade 3	40.55	30.95
Total	**42.70**	**35.74**

These data were analyzed by a two-way Analysis of Variance. The results of this analysis are presented in Table 5,

Table 5: ANOVA Summary Table for the Open-Ended Reading Comprehension Data

Source	SS	df	MS	F	Sig of F
Group	1882.96	1	1882.96	12.09	.001
Grade	651.58	2	325.79	2.09	.163
Group X Grade	230.19	2	115.67	.74	.479
Error	23662.01	152	155.68		

As evident from Table 5, there is a significant effect for group, and nonsignificant effects for grade and for the interaction between group and grade. That is, in general, Program students performed at a significantly higher level than comparison students. Moreover, the difference between the Program and comparison classrooms did not vary as a function of grade level

Although the interaction between group and grade was not significant, it seemed valuable to analyze the difference between Program and comparison students at each grade level. To do this, separate sample t-tests were computed comparing the two groups of students at the first, second and third grades. This analysis showed that the Program students were performing at a significantly higher level at the first and third grades (significance level at the first grade = .042; at third grade the significance level is .005). The difference between the groups at the second grade level, although in favor of the Program students, did not attain significance (p = .146).

Although not of primary interest in this evaluation, data were collected by school for the Program and comparison classes. Since the open-ended score from the SAT-9 is available on students in all three grades, this score is the one which has the most meaning. The mean scores for the open-ended section of the SAT-9 are presented in Appendix A by group and school. The mean scores on the composite section of the SAT-9, for those schools which had at least 3 classrooms in each group at the third grade, are presented in Appendix B.

E · Analysis of the Multiple Choice and Composite Reading Scores

As mentioned previously, the multiple choice test from the SAT-9 was available only for third grade students. Since the composite reading score on the SAT-9 is found by combining the scores from the multiple choice and open-ended sections of the test, this composite is therefore also only available for third grade students. The means, standard deviations and t-test results for the multiple choice section of the SAT-9 are presented in Table 6, while similar data for the reading composite score are presented in Table 7

Table 6: Means, Standard Deviations and t-test Results for the Multiple Choice Section of the SAT-9 (3rd grade students)

	Mean	Standard Deviation	t-test	Significance of t-test
Program Classes	40.55	9.52	2.41	.021
Comparison Classes	30.95	16.05	2.41	.021

Table 7: Means, Standard Deviations and t-test Results for the Composite Reading Score of the SAT-9 (3rd grade students)

	Mean	Standard Deviation	t-test	Significance of t-test
Program Classes	40.09	8.08	2.68	.011
Comparison Classes	31.88	11.86	2.68	.011

As evident in Tables 6 and 7, Program students in the third grade are achieving at significantly higher levels than comparison students on both the multiple choice section of the SAT-9 as well as on the reading composite score.

F · A Further Analysis of Group Differences for First Grade Students

As an additional way to demonstrate Program effects, the data for the first grade students (open-ended SAT-9) were analyzed by dividing the students into the four classifications used by the School District to report competency results. These four classifications are: Below Basic; Basic; Proficient; and Advanced. The percentages of Program and comparison students at each of these levels are presented in Table 8.

Table 8: Percentages of First Grade Students in Program and Comparison Groups at Four Levels of Performance

	Program Students	Comparison Students	Significance of Difference
Below Basic	34.94%	51.53%	.024
Basic	29.69%	31.31%	.709
Proficient	29.28%	14.91%	.007
Advanced	6.08%	2.21%	.068

It is evident from Table 8 that the Program students are performing at significantly higher levels as compared to the non-Program students. Specifically, the percentage of Program students is significantly smaller in the lowest competency group (Below Basic) and significantly higher in the next to highest competency level (Proficient). The difference at the highest level (Advanced), while not quite statistically significant, favors the Program students.

G · Summary of SAT-9 Results Comparing Program and Comparison Classrooms

Several major findings are evident from the analyses presented comparing the Program and comparison classes and students:

- Overall, the students in the Program classrooms are attaining higher achievement levels in reading as compared to the students in the non-program classrooms. This is evident in the multiple choice, open-ended, and composite scores on the SAT-9.
- When analyzed by competency level, the first grade students in Program classes are attaining higher levels.

Where data are available by grade level it is evident that Program effectiveness is not markedly affected by the grade of the student.

H · Analyses Using the Implementation Scores

As mentioned previously, data were available which indicated how extensively the Program had actually been implemented in each classroom. These data include the percentage of students reading 100 books, the percentage reading 200 and 300 books, and an Implementation Score derived from classroom observation. These data can serve as an additional form of within-Program evaluation since it should be the case that students in classrooms in which the Program has been implemented to a greater extent should be reading at higher levels as compared to students in classrooms in which the Program has been less extensively implemented. Several analyses testing this question are presented in this section.

The first analysis consists of Pearson correlations computed between the SAT-9 reading scores and the implementation scores. These correlations are presented in Table 9.

Table 9: Pearson Correlations Between Implementation Scores and SAT-9 Reading Scores

	Multiple Choice	Open-Ended	Composite
% of Students Attaining 100 Books	.430*	.191	.572*
% of Students Attaining 200 Books	.039	.241*	.251
% of Students Attaining 300 Books	-.028	.237*	.164
Implementation Score	.635†	.237*	.462*

It is evident from Table 9 that the results support the validity of the Program. Specifically, the data demonstrate that classrooms in which the Program has been implemented more strongly have students who are achieving at a higher level This is

* $p < .05$

† $p < .01$

clearest in the correlations with the Implementation Score derived from classroom observation, all of which are significant at the .05 level or greater.

As another way of demonstrating the same effect, the distribution of the Implementation Score was divided into thirds to create high, medium and low implementation groups. The mean score on the SAT-9 open-ended test was then used as the dependent variable to ascertain if there was a difference among these three groups. The mean scores for the three groups are presented in Table 10.

Table 10: Mean SAT-9 Open-Ended Scores for Three Levels of Implementation

Low Implementation	39.24
Medium Implementation	43.49
High Implementation	46.03

A one-way Analysis of Variance was used to analyze the difference between the means of the three groups. This produced a marginally significant effect ($F = 3.43$, $p = .035$). It is evident, therefore, that there is a relationship between the extent to which the Program is implemented in a specific classroom and the reading achievement of students in that classroom.

L · Summary and Conclusions

In general, the data presented in this report support the contention that the "100 Book Challenge Program" is having a significant impact on the reading achievement of students participating in the Program. For the first, second and third grade students used for this evaluation, Program participation has produced a significant increase in reading achievement as evidenced by the SAT-9 test data. Moreover, the implementation data indicate that students in classrooms in which the Program has been more strongly implemented demonstrate higher reading achievement than students in classrooms with lower implementation scores. Overall, therefore, the general trend of the data is supportive of the Program.

Having said this, it should also be mentioned that while the students in the Program are achieving higher reading levels than students not in the Program, the average scores are still below the target value of 50 which is reported as the national average. Perhaps, however, this figure is not a realistic or adequate comparison point to be used for students in this study. It must be remembered that the SAT-9 is normed on a national sample which includes, but which is obviously not limited to, urban students. As a means of comparison, therefore, the average score for the entire School District of Philadelphia might be appropriate. These data are presented in Table 11.

Table 11: Program, Comparison, National and School District Average Scores on the SAT-9 for Various Grade Levels

	School District Average	Program Average	Comparison Average
First Grade: Open-Ended	NA	43.72	33.16
Second Grade: Open-Ended	46.1	43.22	38.49
Third Grade: Multiple Choice	40.6	39.17	32.05
Open-Ended	42.0	40.55	30.95
Composite	48.0	40.09	36.41

It is evident from the data in Table 11 that both the Program and the comparison students are achieving at reading levels consistently below the average scores for the School District of Philadelphia as a whole. On one hand, this result is not positive. On the other hand, a reasonable interpretation of these data is that the Program has been implemented in schools which are, in general, more problematic. That is, the fact that both the Program and the comparison students are scoring at a level slightly below the School District's overall average indicates that the schools participating in the Program are those with the most at-risk students. These data, therefore, should be viewed positively as an indication that the Program has not deliberately chosen high achieving schools as a guarantee of success.

Future research on the Program may want to focus on several specific points:

- As the Program is expanded, and as additional grade levels are added to the data set, it will be important to ascertain whether the Program continues its effectiveness at these grade levels. It will be important in future research and future Program expansion to be vigilant about the impact of the Program on older students.
- Because of the nature of the data used in this evaluation, it was not possible to use a longitudinal research design. That is, it would be interesting to follow students as they progress from grade-to-grade, and to follow this progression as a function of Program participation. What happens, for example, to a student who leaves a Program classroom at one grade level and enters a non-Program classroom the next year? Does the impact of the Program continue or diminish? Future research might wish to explore issues of this kind to ascertain the long-term impact of the Program.

- One of the limitations of the design used in this research is that there is no direct way to know if the Program students are comparable to the comparison students. While the use of non-Program classes from the same school to create a comparison group is a reasonable solution, there still might be differences between the two groups that cannot be controlled using the present design. It would be interesting, for example, to use some form of a pretest-posttest design where initial reading level is assessed prior to Program participation. Since the SAT-9 is given yearly in the District, this type of statistical control is possible for future evaluations since a student's SAT-9 score from the previous year (with the exception of students in first grade) can be used in this way.

In summary, the data presented in this report are supportive of the "100 Book Challenge Program". Future research should attempt to expand and improve on this evaluation to ascertain if Program effects are maintained over time and if the impact of the Program on older students is consistent the results from the earlier grades.

Appendix A: Mean Open-Ended SAT-9 Score by Group and School

School	Program Students: Mean Open-Ended Score	Comparison Students: Mean Open-Ended Score
Alcom	42.36	32.07
Anderson	48.89	42.98
Drew	45.70	44.50
Emlen	42.26	40.01
Harrington	45.87	31.58
Hartranft	37.78	35.63
Lowell	38.98	47.98
Meade	39.98	33.09
Richmond	46.46	41.68
Southwark	40.85	NA
Taylor	34.62	22.15
Whittier	51.24	NA

Appendix B: Mean Composite SAT-9 Score by Group and School—Third Grade Only (For Schools with at Least 3 Classrooms in Each Group)

School	Program Students: Mean Composite Reading Score	Comparison Students: Mean Composite Readme Score
Alcorn	41.80	29.69
Harrington	46.20	28.68
Hartranft	36.02	36.04
Meade	29.81	33.88
Richmond	47.79	44.57
Taylor	33.20	27.74