MONTGOMERY COUNTY PUBLIC SCHOOLS

Evaluation Brief

February 2007 Department of Shared Accountability

The Effects of Studying Skillful Teaching Training Program on Students' Reading and Mathematics Achievement

Shahpar Modarresi Ph.D. and Natalie Wolanin[i]

Background

This brief examines the effectiveness of the Studying Skillful Teaching (SST) training on improving students' academic achievement. The SST teachers included in this study may have taken any of the following Skillful Teacher classes or combination of them: SST1, SST2, Observing and Analyzing Teaching (OAT) 1, or 2. Each course is offered by the Montgomery County Public Schools (MCPS) Office of Organizational Development, and comprises 36 or more hours of instruction. The purpose of the training is to promote a common language about skillful teaching and to expand teachers' knowledge of student learning and effective instruction. The following question is addressed in this brief: Do students of teachers who participated in these trainings perform better in reading and mathematics than those students of teachers who did not take the courses, after controlling for teachers' highly qualified status, as well as students' initial abilities, demographics, and service receipt measures?

Methodology

Outcome Measures: Guskey (2000) argues that multiple measures of student learning are essential in examining the influence of teachers' training on students' academic performance. Therefore, the current evaluation used the following two achievement measures: 1) a criterion-referenced scale score from the Maryland School Assessment (MSA) and 2) a scale score from the computer adaptive Measures of Academic Progress-Reading (MAP-R).

Study Sample: These analyses included students who had Grade 3 classroom teachers for whom it could be determined with confidence had taken one of these courses. See Appendix A for a detailed discussion of the methodology. The teachers took SST and/or OAT classes prior to the 2005–2006 school year.

The sample for the reading MSA analysis included 4,822 students who had taken both the 2005 Comprehensive Test of Basic Skills (CTBS) in Grade 2 and the 2006 reading MSA in Grade 3. The MAP-R sample included 5,143 students who had fall 2005 and spring 2006 MAP-R scores in Grade 3. The mathematics sample included 5,674 students who had taken the 2005 CTBS in Grade 2 and who had taken the 2006 mathematics MSA in Grade 3.

Data Analyses: Both statistical significance tests and effect sizes were used to address the evaluation question. The analysis of covariance (ANCOVA) (Kirk, 1995) was utilized to test significant differences between students' mean scores on the outcome measures (MSA and MAP-R) for those taught by teachers who had taken one of the courses and those teachers who had not. For each outcome measure, the ANCOVA model contained the teachers' highly qualified status; the students' prior performance; demographics; and receipt of Free and Reduced-price Meals System (FARMS), special education, and/or English Language Learner (ELL) services; in addition to a propensity score. The propensity score was divided into five categories and used as a categorical covariate in each of the statistical models employed in this evaluation (Rosenbaum and Rubin, 1983, 1984, 1985). The effect sizes were used to judge the practical significance of the observed differences (American Psychological Association, 2001).

Summary of Major Findings

No statistically significant difference was found on the mathematics MSA scores; furthermore, no statistically significant difference was found on MAP-R. While a statistically significant difference was found on the reading MSA, the effect size identified that the observed difference was not practically significant. Therefore, on average, students of SST trained teachers perform as well as students of teachers not trained in SST on reading or mathematics assessments.

Discussion of Findings

Mathematics Outcome Measure

MSA. The descriptive findings indicated that the average mathematics test scores of the students taught by teachers who took one of the courses were higher than those taught by teachers who had not, (CTBS mean difference=2.06 and MSA mean difference=2.70). After controlling for demographics; receipt of FARMS, special education, and ELL services; prior performance; and highly qualified teacher status; the main effect of the training was not significant. This suggests that on average, there were no statistically significant differences on the Grade 3 mathematics MSA scale scores between students taught by teachers who took the training and those taught by teachers who had not taken the training (Appendix Table B1).

Reading Outcome Measures

MSA. The descriptive findings indicated that the average test scores of students taught by teachers who took one of the courses were higher than those taught by teachers who had not taken one of the courses (CTBS mean difference=4.98 and MSA mean difference=5.79). An ANCOVA was performed to detect significant differences on the reading MSA scores, after controlling for the propensity score and teachers' and students' characteristics. The main effect of the training was significant suggesting, that on average, there were statistically significant differences on Grade 3 reading MSA scale scores between the two groups. However, the effect size of 0.07 indicated that the observed significant difference was not of any practical significance (Appendix Table B2). An effect size of 0.2 is considered small; at least 0.5 is considered medium; while 0.8 or greater is considered large (Cohen, 1988).

MAP-R. The descriptive analysis indicated that the average test scores of students taught by teachers who had taken one of the trainings were higher than those taught by teachers who had not had any training (fall 2005 MAP-R mean difference=1.35 and spring 2006 MAP-R mean difference=1.55). The findings from the ANCOVA model revealed that, on average, there were no significant differences between the two groups on the Grade 3 spring MAP-R scale scores. The main effect of the training was not statistically significant as measured by the 2006 MAP-R scores (Appendix Table B3).

Replication of Results

These analyses were repeated using the matching package in R 2.3.1 (R Development Core Team, 2006). The analyses used models that created matched samples, based on propensity scores, and adjusted for teacher highly qualified status, as well as students' demographic and service receipt measures. The results are consistent with the findings presented in this brief (Appendix C).

References

American Psychological Association. (2001). *Publication manual of the American Psychological Association (5th ed.).* Washington, DC: Author.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* Hillsdale, NJ: Lawrence Earlbaum Associates.

Guskey, T. (2000). *Evaluating Professional Development.* Corwin Press, Thousand Oaks, California.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*. Brooks/Cole Publishing Company, New York.

Rosenbaum, P. R., & Rubin, D. B. (1983). *The central role of the propensity score in observational studies for causal effects.* Biometrika, 70, 41–45.

Rosenbaum, P. R., & Rubin, D. B. (1984). *Reducing bias in observational studies using subclassification on the propensity score.* Journal of the American Statistical Association, 79, 561–524.

Rosenbaum, P. R., & Rubin, D. B. (1985). *Constructing a control group using multivariate matched sampling that incorporates the propensity score.* The American Statistician, 39, 33–38.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin Company, Boston, MA.

[i] The authors would like to thank Dr. Scot McNary and Ms. Suzanne Merchlinsky for their contribution to this study.

APPENDIX A

Methodology

Sample for Analysis

The Online Administrative Student Information System (OASIS), supplemented by the report card file, was used to identify Grade 3 students and their assigned reading and mathematics teachers. The original plan for the outcome evaluation was to analyze mathematics and reading data for the entire population of Grade 3 students throughout MCPS. However, due to data limitations, the final analyses included a large sample of Grade 3 students. The authors used the following three decision rules to select the sample of students for the analyses:

1. Exclude students if the teacher listed by OASIS was not a Grade 3 classroom teacher. If a supporting teacher was listed, it could not be determined with confidence which teacher was the student's primary reading or mathematics teacher.
2. Eliminate students who received Grade 4 or 5 reading or mathematics instruction**.**
3. Exclude those students whose teachers' training status could not be determined with confidence. The training database was missing employee ID's for 13% of participants. There were teachers who could not be identified with confidence as having taken or not taken an SST or OAT course.

The final sample for the analyses included students who had Grade 3 reading and mathematics classroom teachers whose training status could be determined with confidence. Please note that results cannot be generalized to students who received special instruction from supporting teachers or from higher grade-level teachers.

The reading sample for the MSA analysis included 4,822 students who had complete data records, including Grade 2 2005 CTBS and Grade 3 2006 MSA Reading scores. The sample for the MAP-R analysis included the 5,143 students who had fall 2005 and spring 2006 MAP-R reading scores for Grade 3. The mathematics sample for the MSA analysis consisted of 5,674 students who had Grade 2 2005 CTBS and Grade 3 2006 MSA Mathematics scale scores.

Evaluation Design

Due to the fact that students were not randomly assigned to teachers, this evaluation used a nonequivalent control group, pre- and posttest design, a frequently used type of quasi-experimental design (Table B1). The stated design is the most appropriate evaluation design in assessing the effectiveness of any intervention program among the quasi-experimental designs (Shadish, Cook & Campbell, 2002). A problem with this design is that the two groups of students might differ in important ways that may influence their performance. The advantage to this design is that the preexisting differences between groups of students can be accounted for in the statistical analysis.

The Design of the SST Professional Development Evaluation

Group	Pretest		Instructional Delivery		Posttest
Students of Teachers With Training	$O_{1a \text{ and } b}$	=>	X	=>	$O_{2a \text{ and } b}$
Students of Teachers Without Training	$O_{1a \text{ and } b}$	=>	C	=>	$O_{2a \text{ and } b}$

O_{1a} – Spring 2005 CTBS scores (reading and mathematics)
O_{1b} – Fall 2005 MAP-R reading scale scores
X – The instructional delivery by teachers who had taken SST courses (treatment group)
C – The instructional delivery by teachers who had not taken SST courses (comparison group)
O_{2a} – Spring 2006 MSA scale scores (reading and mathematics)
O_{2b} – Spring 2006 MAP-R reading scores

Analysis Procedures

Propensity scores (based on teachers' highly qualified status, as well as students' pretest scores, demographics, and service receipt measures) were computed using Logistic Regression and saved as a new variable in the data sets. The propensity score is a predicted probability of receiving the intervention (in this case, of being taught by a teacher who had the training), based on students' characteristics and teachers' highly qualified status. The variable was further divided into five categories and used as a categorical covariate in each of the ANCOVA models in this study. The demographic and initial abilities of students, as well as the teachers' highly qualified status, also were included in the ANCOVA models to reduce the residual variability of the outcomes (MSA and MAP-R test scores). To test for non-parallelism or interaction (homogeneity of regression slopes), the product term between the pretest scores and the group variable was included in each of the ANCOVA models. The evaluation of the students' reading and mathematics performance in Grade 3 was conducted by constructing the following three models:

> Model I. The dependent variable for this model was the spring 2006 mathematics MSA scale scores. The independent variable was a dummy variable created to represent the status of the students' experience. The control variables or covariates included race/ethnicity; receipt of FARMS, special education, and/or ELL services; and highly qualified teacher status; plus a propensity score. The pretests for this cohort were the spring 2005 CTBS mathematics scale scores. The correlation coefficient of Grade 2 CTBS mathematics scores with Grade 3 mathematics MSA was significant ($r=0.69$; $p<0.05$).
>
> Model II. The dependent variable for this model was the spring 2006 reading MSA test scores. The independent variable was a dummy variable created to represent the status of the students' experience. The control variables or covariates included race/ethnicity; receipt of FARMS, special education, and/or ELL services; and highly qualified teacher status; plus a propensity score. The pretests for this cohort were the spring 2005 CTBS reading scale scores. The correlation coefficient of the Grade 2 CTBS reading scores with Grade 3 reading MSA was significant ($r=0.67$; $p<0.05$).
>
> Model III. The dependent variable or the outcome measure for this model was the spring 2006 MAP-R reading test scores. The same independent and control variables (or covariates) as the one indicated in the previous models were used. The pretests for this cohort were the fall 2005 MAP-R scale scores. The correlation coefficient between fall 2005 MAP-R and spring 2006 MAP-R was significant ($r=0.83$; $p<0.05$).

Reliance solely on the significance test may lead one to accept an effect of trivial magnitude. Test statistics and their p-values are greatly affected by the study's sample size. Therefore, the effect sizes of the differences were investigated to determine the magnitude of effects associated with mean differences. The following formula was used to calculate the effect size in this evaluation: effect size = $(M_t - M_c)/SD$. The M_t and M_c are adjusted group means for students of teachers who were trained and those who were not, respectively, and SD is the standard deviation of the pooled posttest scores.

APPENDIX B

Table B1
Adjusted Means, Mean Difference, and Effect Size for the 2006 Mathematics MSA

	Adjusted Means		Treatment Effect	
Outcome Measure	SST Students *N*=2487	Non-SST Students *N*=3187	Mean Difference	Effect Size
Mathematics MSA	423.7	421.9	1.8	0.04

(F=3.05; P>0.05)

Table B2
Adjusted Means, Mean Difference, and Effect Size for the 2006 Reading MSA

	Adjusted Means		Treatment Effect	
Outcome Measure	SST Students *N*=2051	Non-SST Students *N*=2771	Mean Difference	Effect Size
Reading MSA	419.5	416.8	2.7	0.07

(F=7.56; P<0.05)

Table B3
Adjusted Mean, Mean Difference, and Effect Size for the 2006 MAP-R

	Adjusted Means		Treatment Effect	
Outcome Measure	SST Students *N*=2200	Non-SST Students *N*=2943	Mean Difference	Effect Size
MAP-R Reading	202.2	201.7	0.44	0.03

(F=0.72; P>0.05)

APPENDIX C

Replication of Analyses using the Matching package in R 2.3.1 (R Development Core Team, 2006)

Scot McNary, Ph.D.

The mean differences between students taught by teachers who had taken one of the trainings and those taught by teachers who had not, are calculated based on matched samples (matched on propensity scores), with covariates included (race/ethnicity; receipt of FARMS, special education, and ELL services; highly qualified teacher status; and prior performance).

Effect sizes were calculated from a matched sample design, which is different than the independent groups design. The matched sample design can be thought of as testing, whether or not the (adjusted) mean difference score between the matched pairs is significantly different from zero. This was treated as a one sample hypothesis test for the mean, with the observed statistic equal to the mean difference score, divided by the standard error of the difference score. In the one sample hypothesis test for the mean, the effect size is calculated by t/√df (t divided by the square root of the degrees of freedom for the test). The degrees of freedom for these tests are N - # covariates - 1.

The findings (Table C1) suggest that the training did not raise the performance of Grade 3 students as measured by MSA and MAP-R.

Table C1
Estimates for Grade 3 Analyses Using MSA and MAP-R

	Difference	Standard Error	t	Effect Sizes	Degrees of Freedom	P
Grade 3						
Reading MSA	1.45	0.79	1.83	0.03	4803	0.07
MAP-R	0.26	0.25	1.05	0.004	5121	0.29
Mathematics MSA	1.36	0.76	1.79	0.02	5649	0.07