MONTGOMERY COUNTY PUBLIC SCHOOLS

Evaluation Brief

February 2007 Department of Shared Accountability

The Effects of Studying Skillful Teaching Training Program on Students' Algebra Achievement

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

Background

This brief summarizes the findings of an evaluation conducted to examine the effectiveness of the Studying Skillful Teaching (SST) or Observing and Analyzing Teaching (OAT) training in improving students' achievement in the Algebra High School Assessment (HSA). The Montgomery County Public Schools (MCPS) SST teachers included in this study may have taken any of the following Skillful Teacher classes or combination of them (SST1, SST2, OAT1, or OAT2). Each of these classes provides at least 36 hours of instruction. The SST training is based on the framework of instructional parameters (i.e., management, instructional strategies, motivation, and curriculum planning) and a foundation of essential beliefs about student learning (Saphier & Gower, 1997). OAT is based on the same principles as the SST courses. This brief addresses the following question: Do students of trained teachers perform better on the Algebra HSA than students of teachers who had not had the training after controlling for teachers' highly qualified status, as well as students' initial abilities, demographics, and service receipt measures?

Methodology

Design. A nonrandomized comparison group pre- and posttest design (Isaac & Michael, 1995) was used to assess the effectiveness of the training program on students' performance on the Algebra HSA. The pretest was the mathematics Maryland School Assessment (MSA) and the posttest was the Algebra HSA .

Study Sample. The analyses included secondary students who had the same Algebra 1 teacher throughout the 2005–2006 school year, and it could be determined with confidence that the teacher had taken SST and/or OAT classes prior to the 2005–2006 school year. Appendix A contains a detailed discussion of the methodology. Students may have taken Algebra 1 in several grades, so analyses were conducted by grade level, (i.e., Grades 7, 8, or 9). Students taking Algebra 1 in Grade 10 were excluded from the analyses due to the small sample size (N=113).

The sample for the Grade 7 analysis included 1,230 students who had Grade 6 mathematics MSA and Grade 7 Algebra HSA scores. The Grade 8 sample included 2,697 students who had Grade 7 mathematics MSA and Grade 8 Algebra HSA scores. The Grade 9 sample consisted of 882 students who had Grade 8 mathematics MSA and Grade 9 Algebra HSA scores.

Data Analyses. The analysis of covariance (ANCOVA) was used to test significant differences between students' mean scale scores on the Algebra HSA (Kirk, 1995). The ANCOVA model contained the teachers' highly qualified status, as well as 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 to control for confounding due to preexisting differences between the two groups (Rosenbaum & Rubin, 1983). Effect sizes were used to estimate the magnitude of SST training effects (American Psychological Association, 2001).

Summary of Major Findings

No statistically significant differences were found for performance on the Algebra HSA for students of teachers who had the training compared with students of teachers who had not had the training. A summary of findings for each of the grade levels is discussed below.

Discussion of Findings

Appendix B contains the adjusted means and effect sizes for the three cohorts included in this study.

Grade 7 Sample. The descriptive findings indicated that the average test scores of students of teachers who had the training were higher than those of teachers who did not on both the pre- (mathematics MSA mean difference=7.0) and posttests (Algebra HSA mean difference=4.1). An ANCOVA was performed to detect significant differences between the two groups after controlling for propensity score, as well as teachers' highly qualified status and students' service receipt measures (Appendix A). The main effect of the training was not significant, indicating that on average, there was not a statistically significant difference between the two groups on the 2006 Grade 7 Algebra HSA .

Grade 8 Sample. The descriptive findings indicated that the average test scores of the students of trained teachers were lower than students of non-trained teachers on pre- (mathematics MSA mean difference=-1.0) but higher on posttests (Algebra HSA mean difference=1.9). The main effect of the training was not significant, suggesting that on average, there was not a statistically significant difference on the 2006 Grade 8 Algebra HSA scores between the two groups.

Grade 9 Sample. The descriptive findings indicated that the average test scores of students of teachers who had the training were lower than those of teachers who did not on both on both pre- (mathematics MSA mean difference=-1.7) and posttests (Algebra HSA mean difference=-1.8). The main effect of the training was not significant, indicating that on average, the two groups did not score significantly different on the Algebra HSA .

Replication of Results

The above analyses were repeated by a second researcher using an independent model and software (matching package in R 2.3.1, R Development Core Team, 2006) employing the same data set. The analyses used models that created matched samples based on propensity scores and adjusted for teachers' highly qualified status, as well as students' demographic and service receipt measures. The results are consistent with the findings presented in this brief, with one exception (Appendix C). The results for Grade 8 show statistically significant differences between the two groups of students. However, the practical difference, as measured by an effect size, was too small for a practical significance (0.09) and was the same as that calculated in the original analysis.

Limitations of the Findings

Although the findings obtained from this study were based on sound evaluation design, as well as appropriate analyses, it should be noted that causality should not be inferred from the current study.

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.

Isaac, S. & Michael, W. (1995). *Handbook in research and evaluation, (3rd Ed.).* EDITS/Educational and Industrial Testing Service, CA.

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.

Saphier, J., and R. Gower (1997). *The Skillful Teacher: Building Your Teaching Skills.* Acton, Massachusetts, Research for Better Teaching.

[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

MCPS report card data were used to retrieve information regarding Algebra 1 students and their assigned Algebra 1 teachers. The original plan for the SST outcome evaluation was to analyze Algebra HSA scores for all students. However, during the file development, the authors used two decision rules to select the sample of Algebra 1 students for the analyses.

1. Exclude students if their teachers for Semester 1 and Semester 2 differed. The goal was to assure the same teacher instructed students throughout the 2005–2006 school year, and to eliminate any differences that might have occurred with two different teachers. A common practice for high school schedules is that a student may have a different teacher for each semester.

2. Exclude students whose teachers' training status (SST or OAT) could not be determined with confidence. The teachers' training database lacked employee ID's for 13% of its participants. As a result, some teachers were excluded from the study.

It also should be noted that because of one method used to identify many of the teachers, a higher number of male teachers than female teachers exists in the non-trained teacher category.

Evaluation Design

This evaluation used a nonequivalent control group, pre- and posttest design, a frequently used type of quasi-experimental design.

The Design of the SST Professional Development Evaluation

Group	Pretest	Instructional Delivery	Posttest
Students of Teachers With Training	O_1 =>	X =>	O_2
Students of Teachers Without Training	O_1 =>	C =>	O_2

O_1 – Spring 2005 mathematics MSA scale scores (Grades 6, 7, and 8)
X – The Algebra 1 instructional delivery by trained teachers
C – The Algebra 1 instructional delivery by non-trained teachers
O_2 – Spring 2006 Algebra HSA scale scores

Analysis Procedures

Both statistical significance tests and effect size measures were used to assess program effects in this evaluation. The ANCOVA (Kirk, 1995) was used to test for achievement differences between the two groups' Algebra HSA scale scores, after controlling for students' prior achievement and characteristics, as well as teachers' highly qualified status. Propensity scores (based on students' pretest scores, demographics, and service receipt measures, as well as teachers' highly qualified status) were computed using a logistic regression model. The propensity score was divided into five categories and used as a categorical covariate in each of the statistical models (Rosenbaum & Rubin, 1983, 1984, 1985). The students' characteristics, as well as the teachers' highly qualified status, also were included in the ANCOVA models to reduce the residual variability of the outcomes (Algebra HSA test scores). To test for non-parallelism or interaction (homogeneity of regression slopes), the product term between pretest scores and group variable was included in each of the ANCOVA models. The evaluation of the students of SST/OAT trained teachers on Algebra 1 performance was conducted by constructing the following three models:

> *Grade 7 Model*. The dependent variable for this model was the spring 2006 Algebra HSA scale scores. The independent variable consisted of two categories: a) students of trained teachers and b) students of non-trained teachers. The control variables included receipt of FARMS, special education, and/or ELL services; as well as highly qualified teacher status and the propensity score. The pretests for this cohort were the spring 2005 Grade 6 mathematics MSA scale scores. The correlation coefficient of Grade 6 mathematics MSA scores with the spring 2006 Algebra HSA was significant ($r=0.52$; $p<0.001$).
>
> *Grade 8 Model*. The dependent variable for this model was the spring 2006 Algebra HSA scale scores. The same independent and control variables (or covariates) as the one indicated in the Grade 7 model also were used in the Grade 8 model. The pretest for this model was the spring 2005 Grade 7 mathematics MSA scale scores. The correlation coefficient of Grade 7 mathematics MSA scores with the spring 2006 Algebra HSA was significant ($r=0.63$; $p<0.001$).
>
> *Grade 9 Model*. The dependent variable for this model was the spring 2006 Algebra HSA scale scores. The same independent and covariates used in the previous models were employed in this model. The pretest for this cohort was the spring 2005 Grade 8 mathematics MSA scale scores. The correlation coefficient of the Grade 8 mathematics MSA scores with the spring 2006 Algebra HSA was significant ($r=0.60$; $p<0.001$).

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 trained and non-trained teachers respectively, and SD is the standard deviation of the pooled posttest scores. Based on the Cohen's convention (1988), an effect size of 0.2 is considered small, an effect size of 0.5 is considered medium, and an effect size of 0.8 or greater may be considered large.

Since random assignment procedures were not followed, causal conclusions about the impacts of the training on the Algebra 1 achievement of students, as measured by the Algebra HSA may not be inferred from the analyses. However, the use of both propensity score covariates and control variables in the current evaluation improved the internal validity of the study design.

APPENDIX B

Table B1
Adjusted Means, Mean Difference, and Effect Size for the 2006 Algebra HSA Test Scores
(Grade 7 Sample)

	Adjusted Mean		Treatment Effect	
Outcome Measure	SST Students *N*=872	Non-SST Students *N*=358	Mean Difference	Effect Size
Spring 2006 Algebra HSA	467.8	467.4	0.41	0.02

(F=0.375; P>0.05)

Table B2
Adjusted Means, Mean Difference, and Effect Size for the 2006 Algebra HSA Test Scores
(Grade 8 Sample)

	Adjusted Mean		Treatment Effect	
Outcome Measure	SST Students *N*=1687	Non-SST students *N*=1010	Mean Difference	Effect Size
Spring 2006 Algebra HSA	443.5	441.5	1.91	0.09

(F=0.08; P>0.05)

Table B3
Adjusted Means, Mean Difference, and Effect Size for the 2006 Algebra HSA Test Scores
(Grade 9 Sample)

	Adjusted Mean		Treatment Effect	
Outcome Measure	SST Students *N*=308	Non-SST Students *N*=574	Mean Difference	Effect Size
Spring 2006 Algebra HSA	411.1	410.8	0.27	0.01

(F=0.03; P>0.05)

APPENDIX C

Analyses Using The Matching Package in R 2.3.1 (R Development Core Team, 2006)

Scot McNary, Ph.D.

The analyses addresses the following question: Do students of trained teachers perform better on the Algebra HSA than those students of teachers who have not had the training after controlling for teachers' highly qualified status, as well as students' initial abilities, demographics, and service receipt measures?

The mean differences between the two groups of students are calculated based on matched samples (matched on propensity scores), with covariates included, as well as race/ethnicity, receipt of FARMS, special education, and/or ELL services; highly qualified teacher status; and prior performance. The matching package does more than include the propensity score as a covariate; it also creates a matched sample and uses covariates.

Effect sizes are 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 the (adjusted) mean difference score between the matched pairs is significantly different from zero. This is treated as a one sample hypothesis test for the mean, with the observed t 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 $t/\sqrt{df}$ (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 indicate, that on average, students of trained teachers perform as well as matched students of non-trained teachers, as measured by their Algebra HSA test scores (see Table C1).

Table C1
Estimates for Algebra HSA Analyses

	Difference	Standard Error	t	Effect Size	Degrees of Freedom	P
Algebra HSA						
Grade 7	0.70	1.49	0.47	0.01	1092	0.64
Grade 8	3.53	0.86	4.13	0.09	2250	<0.001
Grade 9	1.21	1.76	0.69	0.03	745	0.49