Using the Multiple-Matched-Sample and Statistical Controls to Examine the Effects of Magnet School Programs on the Reading and Mathematics Performance of Students

By

Yu, N. Yang[1], Yuan H. Li & Leroy J. Tompkins

Prince George's County Public Schools, Maryland

Shahpar Modarresi

Montgomery County Public Schools, Maryland

Paper was presented at the annual meeting of the American Educational Research Association, Montreal, Canada, April, 2005.

[1] The views are those of the authors, and no official support by the Prince George's County Public Schools is intended.

Using the Multiple-Matched-Sample and Statistical Controls to Examine the Effects of Magnet School Programs on the Reading and Mathematics Performance of Students

Abstract

This summative evaluation of magnet programs employed a quasi-experimental design to investigate whether or not students enrolled in magnet programs gained any achievement advantage over students who were not enrolled in a magnet program. Researchers used *Zero-One Linear Programming* to draw multiple sets of matched samples from the non-magnet student population to serve as multiple control groups for the research effort. Whenever a matched sample was generated, the analysis of covariance was subsequently used to control for the effects of the possible pretest score difference between the two groups on the outcome variable (posttest score on reading or mathematics). The mean of the effect size values, across 200 matched-sample analyses, was used to investigate the effects of the magnet program treatment on the reading and mathematics performance of magnet students.

The results for the elementary school magnet programs showed that when students' demographics and initial abilities were accounted for, only the French immersion program had a likely positive impact on student's reading or mathematics performance. Nevertheless, the rest of the magnet programs had minimal, if any, positive effect to promote higher reading and mathematics test scores of magnet students.

Key Words: Magnet Program Evaluation, Linear Programming, Matched Sample, Quasi-Experimental Design, Optimization, Experimental Design, Program Evaluation

I. Introduction

A. Magnet School Programs Overview

The purpose that magnet programs serve in most public school systems has evolved over time. Initially, magnet programs were set up to serve as an alternative to mandatory student reassignment (busing) in the context of federal school desegregation litigation. The programs were offered as an alternative curriculum – albeit, not necessarily superior – to the general public school curriculum. It was thought that the attraction of alternative programs would be sufficient to encourage parents to send their children into racially diverse education environments.

These alternative academic programs required a resource investment that was over and above the resource level required for regular comprehensive schools. For the past two decades, the Maryland state legislature has supported this additional investment with an annual magnet grant. Recently, however, federal oversight in a school system's desegregation case has begun to phase out, the purpose of magnet programs in the eyes of public officials and the community at-large began to change. No longer are magnet programs viewed as a desegregation tool, rather they are viewed as alternative and superior academic programs. As such, if magnet programs are to be perceived as successful in the eyes of the general public, students who attend these programs are expected to outperform their non-magnet peers on standardized academic achievement tests. And if magnet schools cannot produce higher performing students, many would argue that the programs are not worth the added cost of operation.

B. Background of Methodology Used for the Summative Evaluation

In order to determine the impact of magnet programs on student academic achievement, summative evaluations have often been undertaken. A review of magnet program evaluation literature found several studies that examined student outcomes in magnet schools. Among them, a value-added study (Gamoran, 1996) is a valuable reference to be studied while conducting such type of studies. That study compared a subset of 48 magnet schools to 213 regular high schools in the 1988 National Educational Longitudinal Survey and found that magnet school students outperformed their peers attending regular schools in social studies, science, and reading. Another value-added study conducted by Adcock and Phillips (2000), however, found that magnet program students did not perform as well as their non-magnet counterparts.

In order to better determine the efficacy of magnet programs, a purely statistical modeling (e.g., Hierarchical linear modeling, HLM, Bryk & Raudenbush, [1992]; Analysis of covariance, ANCOVA, Kirk, 1995) was used in both value-added studies, indicated above, to account for student's characteristics (e.g., sex, race, pretest scores, etc.) as well as school context (e.g., percent of minority students, percent of poverty students, etc.). However, as pointed out by Rubin, Stuart and Zanutto (2004), comparing results obtained from treated (e.g., magnet programs) and whole control groups (e.g., non-magnet population) with very different distributions of background covariates will heavily rely on modeling (e.g., HLM) assumptions that cannot be tested and extreme extrapolation. Reliable causal inferences may thus not be drawn. For example, Rubin et. al (2004) mentioned that the values of "percent minority" and "percent in poverty" may differ widely in some schools, and this situation will cause the estimated program effects that have been adjusted for such covariates to be extremely sensitive to these statistical modeling assumptions (e.g., parallel slopes). If the assumptions are seriously

violated, those estimated program effects, as a result of using extreme extrapolation, would be seriously misleading.

In comparing the program (or school) effects among magnet programs themselves, seeking an appropriate matched sample to be compared with for each magnet program (or school) is expected to be a necessary step before any statistical modeling is performed in assessing program effects. However, how to create an appropriate matched sample for each magnet program is another challenging issue. Diverse methods can be used to serve this purpose (for literature review, see Shadish, Cook & Campbell, 2002). Among them, using the logistic regression modeling to compute the probability (or propensity score) as a criterion for selecting students as members of a matched sample is one of the promising approaches to address this issue. However, the value of the propensity score is dependent on the selections of logistic regression models (e.g., whether or not including the interaction, nonlinear terms, etc.). Also, if the assumptions made for the logistic regression are not met, and/or if the sample size used for statistical modeling is not large enough, using the propensity score as a criterion for selecting a matched sample might not be as meaningful as researchers anticipated.

Furthermore, the weighting for each covariate, that is then used for computing the propensity score, depends totally on the degree of each covariate's relation to the treatment assignment (received or not received treatment). This procedure is not appropriate for the *non-randomized comparison group pretest-posttest design*, in which the pretest score is usually highly correlated with the outcome measure rather than the treatment assignment. This scenario will cause the pretest-score covariate to be less important than it should be when the propensity score is used to select a matched sample.

It seems to be appropriate to state that the *non-randomized comparison group pretest-posttest design* (refer to Shadish, Cook & Campbell [2002], p. 136) is an appropriate evaluation design in assessing the efficacy of any program among the quasi-experimental designs. Because of the limitations of the propensity score in this design, the current study utilized a *Zero-One Linear Programming* approach to create a matched sample as a control group for the quasi-experimental designs (for technical details, refer to Appendix C, Li, Yang, Tompkins & Modarresi 2005). Several studies (e.g., Li and Schafer [2005a], Li and Schafer [2005b], Theunissen [1985 and 1986], and van der Linden and Boekkooi-Timminga [1989]) have successfully utilized this technique in the area of educational measurement for creating multiple tests with similar characteristics (e.g., item difficulties, test information, and test specifications). Compared to the existing propensity score matching method, this matching method does not require the choice of a statistical model, often used for the computation of propensity score. This prevent any negative consequence that might occur when any assumptions made for the selected model is seriously violated. Moreover, this matching procedure can handle the covariate of the pretest score more appropriately and is very efficient in matching as many demographic and initial ability (or pretests) variables as the researcher desires. The identical distribution of different types of students (Male/White/Poverty, Female/White/Poverty, Male/Asian/Poverty, etc.) between the experimental and matched samples is a promising feature that can hardly be found in pre-existing matching procedures in the literature.

If the measurement error of the pretest-score mean of the magnet program group is ignored, the method introduced above will generate a *unique* matched sample once the criteria for attempting to create two similar groups are determined. For large sample sizes, like those found in several magnet programs (e.g., Academic center program), the assumption of non-measurement-error should be appropriate. Nevertheless, this assumption is improper for the programs with small sample sizes (e.g., Music program). To increase the confidence level of creating an appropriate matched sample as similar to the treatment group as we could obtain, such non-measurement-error is not necessary to be presumed by allowing the pretest-score mean to be contaminated with a "reasonable" measurement error. Appendix C delineates the detailed steps used to generate such "reasonable" measurement error.

Allowing the addition of error into the average of the pretest score of the treatment group during the process of matching procedures may generate multiple similar matched samples due to the fact that multiple "reasonable" measurement errors may exist. This condition allows the reading and mathematics performance of each magnet program group to have multiple matched samples to compare with. Afterwards, the mean of the effect size measures (to be defined later), taking the average of the effect size across replicated comparisons, can then be used to assess the efficacy of any program. This method enables researchers to have more confidence in deciding whether a program (especially for a programs with a small sample size) is effective or not than a single-matched –sample control method does. Accordingly, the multiple-matched-sample control, accommodated with the ANCOVA statistical control, was used in this study for each magnet program outcome evaluation.

The goal of obtaining multiple sets of matched samples as multiple sets of control groups was to create the conditions similar to replicated-randomized-assignment experiments as closely as possible. In this evaluation, the treatment and control groups were matched without using the observed outcome variable (or posttest), thus preventing us from "intentionally" manipulating any sets of matched samples to obtain a desired result and also protecting from such claims by researchers The ability of a matched sample procedure to reveal the extent to which treated and matched groups have similar types of students in similar educational settings is "an important diagnostic tool to identify whether the data can support [possible] causal comparisons between these two groups" (Rubin et. al. 2004).

C. Evaluation Purpose

This study attempted to determine whether there is the probability of a program effect from the various magnet programs that operate in a public school system. Its primary purpose is to determine whether or not attending magnet programs leads to higher student achievement. This purpose can be achieved by comparing the academic performance of students from each magnet program to multiple sets of matched samples of non-magnet peers in both reading and mathematics. A secondary, but related purpose is to determine which programs contributed the most in promoting higher academic achievement. The following two questions directed the design of this evaluation:

1. Do students in a magnet program perform better in reading and mathematics than multiple sets of matched samples of non-magnet students after controlling for students' demographics (viz., poverty, race, and gender) and their initial abilities?

2. Among magnet programs, which were the most effective in improving students' academic performance in reading and mathematics after controlling for students' demographics and their initial abilities?

II. Methodology

A. Evaluation Design

A *Non-Randomized Comparison Group Pretest-Posttest Quasi-Experimental Design* (p.76, Isaac & Michael, 1995) was used to assess magnet program effects on students' reading and mathematics achievement. The magnet student population for each program type served as the experimental groups. Multiple sets of matched samples were drawn from the non-magnet student population (to be discussed later) to serve as multiple control groups for each program evaluation. Figure 1 (below) illustrates the evaluation design.

	Pretest		Treatment		Posttest
Quasi-Experimental Group *(Magnet Student Population)*	O_1	=>	X	=>	O_2
Control Group *(Multiple Matched Samples drawn from Non-Magnet Student Population)*	O_1	=>	C	=>	O_2

O_1 – SY 2002 CTBS results for fourth graders in Mathematics and Reading
X – Magnet Program treatment
C – Regular (Comprehensive) program treatment
O_2 – SY 2003 MSA results for fifth graders in Mathematics and Reading

Figure 1: The Evaluation Design of the Elementary School Magnet School Programs

B. Measures of Student Performance

The elementary school cohort includes fifth grade students enrolled in elementary schools during the 2002-2003 school year. The dependent variables for this evaluation were the Spring 2003 Maryland School Assessment (MSA) reading and mathematics test scores. The pretests for this cohort were the Spring 2002 Comprehensive Tests of Basic Skills (CTBS) reading and mathematics scores. Any students receiving special education or ESOL services were excluded in this analysis.

C. Matching Control

A statistical technique, *Zero-One Linear Programming* (LINDO Systems, Inc., 2003), was used to create a *Matched Sample*. Multiple sets of matched samples for each program evaluation were drawn from the non-magnet program student population with the following constraints:

a) the average pretest score of the non-magnet matched sample was close to the average pretest score of the Magnet Program sample (See Table 1)
b) the matched sample had the same demographic characteristics (viz., race, gender, and poverty status — free/reduced or paid lunch) as the magnet sample.

For example, as shown in Table 1, there were 468 and 6,435 fifth graders in the Academic Center Magnet Program and comprehensive (or Non-Magnet) programs, respectively. Students were grouped by combinations of race, gender, and poverty status, i.e., 20 types of students were classified and listed in the first column of Table 1.The frequencies of those 20 types of students are also shown in Table 1 for both Magnet and Non-Magnet groups.

The pretest means of CTBS reading were 642.68 and 642.44 for the Academic Center group and respective matched group. This matched sample could be one of multiple sets of matched sample that was used for the evaluation of reading performance of the Academic Center Magnet Students. No statistically significant difference was found for both means. Also, constraint b made the number of students for each selected matching variable identical between the two groups. Besides that, as seen in Table 1, the distribution of students in the combination of those selected matching variables in both groups was identical. The latter feature combined with the first constraint make the matched sample generated by this matching method as similar to the Academic Center group as we could obtain.

Table 1:
Frequency Distributions and Average Pretest Scores for the Academic Center and Its Matched Sample

Types of Students	Magnet Program Frequency	Matched Sample Frequency	Non-Magnet Students Frequency
1. American Indian, male, non-poverty	1	1	5
2. American Indian, male, poverty	0	0	7
3. American Indian, female, non-poverty	3	3	9
4. American Indian, female, poverty	0	0	8
5. Asian, male, non-poverty	3	3	53
6. Asian, male, poverty	3	3	34
7. Asian, female, non-poverty	7	7	55
8. Asian, female, poverty	2	2	32
9. African American, male, non-poverty	92	92	1,098
10. African American, male, poverty	80	80	1,399
11. African American, female, non-poverty	100	100	1,121
12. African American, female, poverty	103	103	1,548
13. White, male, non-poverty	19	19	215
14. White, male, poverty	0	0	46

15. White, female, non-poverty	19	19	219
16. White, female, poverty	4	4	44
17. Hispanic, male, non-poverty	9	9	35
18. Hispanic, male, poverty	15	15	210
19. Hispanic, female, non-poverty	2	2	46
20. Hispanic, female, poverty	6	6	251
Total N	**468**	**468**	**6,435**
Average Pretest Score	**642.68**	**642.44**	
Difference of Pretest Score	**0.24***		

* P = .986

Appendix C delineates the detailed steps used to draw a matched sample or multiple sets of matched samples from the non-magnet student population employing *Zero-One Linear Programming*. If both magnet and non-magnet groups have more overlapping distributions on those matching variables, then the matched sample could be adequately obtained without the need of selecting members from extreme tails of the distributions. For example, the non-magnet population might have more overlapping distributions if such population is composed of more members who are eligible for the magnet program, but they are not placed in the magnet program due to some circumstances (e.g., schedule conflict, no intention to attend, etc.). In contrast, the magnet population might have less overlapping distributions if such population is only composed of members who are not eligible at all. When the later scenario occurs, examination of the overlap of the two distributions will help alert researchers to the possibility of a regression effect among the matches (Shadish, Cook, & Campbell, 2002, p 121).

D. Statistical Control

After the matching procedure, a small pretest score difference between the magnet sample and its matched sample remained. The analysis of covariance (ANCOVA) was used to control for the effects of the pretest score difference (e.g., on the CTBS Reading Test). Technically, ANCOVA resulted in *adjusted posttest means* for both groups under the constraint of two groups' pretest means being equal. When the ANCOVA analysis was incorporated with the matched sample design, the impact of possible violation of ANCOVA statistical assumptions on the estimate of the adjusted posttest means is expected to be minimal, as explained in the Appendix A that delineates the assumptions underlying the use of ANCOVA.

E. Evaluation Criterion: The Average of Multiple Effect Size Measures

The adjusted means estimated by the model were computed for a magnet program and its matched sample. The *value-added score* (or non-standardized effect size) was then obtained by computing the Adjusted $\text{Mean}_{\text{Magnet Program}}$ minus Adjusted $\text{Mean}_{\text{MatchedSample}}$. Since the magnitude of the *Value-Added Score* is primarily dependent on metric of the posttest score, a standardized effect size (called ES), with a promising feature of scale invariant or metric-free, was subsequently computed, as illustrated in equation 1.

$$ES = \frac{Value\ -Added\ Score}{SD_{pooled\ Posttest\ Score}} \quad (1)$$

As seen in Equation 1, the standardized ES is defined as the *Value-Added Score* divided by the standard deviation (SD) of the pooled posttest scores (Thompson, 2002). Due to metric-free feature of the standardized ES, it can be used to compare the treatment effects among multiple comparisons. Also, its mean, across multiple matched-sample analyses for each program evaluation, is meaningful. In this evaluation, two hundred replicated matched samples were created and two hundred ES measures were then obtained. The mean of those 200 ESs was primarily used to assess each program's effectiveness.

The standardized ES can also be thought of as the percentile rank (PR) standing of the magnet program sample mean when it compares with the distribution of the matched-sample test scores. Cohen (1988) suggested that a 0.2 ES may be labeled as small; an ES of at least 0.5 as medium; while an ES of 0.8 or greater may be considered large. Accordingly, the average effect size of 0.2 is required to show efficacy of the magnet program in this study. The interpretation of the ES is provided in Appendix B.

III. Results

A. Findings of Academic Achievement of All Magnet Programs

At the time this evaluation was conducted, there were seven magnet programs (for the complete listing of current magnet programs, see Table 2 shown below) at the elementary school level. Because the nature of the talented and gifted program is unique, this program was excluded in the current analysis.

After students' demographics and initial abilities were accounted for across both magnet programs and matched samples of students, the summary results of effect size analysis in reading and mathematics for each magnet program were listed in Table 3.

The mean of the effect size measures is a key index to gauge the program effectiveness. This index suggested that the French Immersion had a positive impact on students' reading (ES=0.23) as well as mathematics (ES=0.32) performance. The mean of 200 effect size values seems to indicate that the rest of the magnet programs had minimal, if any, program effect for grade four students enrolled in this program on their succeeding grade five MSA reading or mathematics performance because their respective average ES values were less than 0.2. Some values were negative. It is important to note that the effect size value for the Music & Technology program must be interpreted with extreme caution because the program was only implemented for half of a year. As such, the students only received about six months of program treatment. In addition, the sample size of this program is not very large (N=32), so the effect size of this program might not be as reliable as those obtained from larger sample sizes. The larger standard deviation of the effect size for the Music and Technology program is another index to support this concern.

The distribution of effect size measures for each magnet program was plotted on Figures 2 thru 15. For evaluating each magnet program, its respective distributional plot of effect size measures provides the richest information for decision makers. Such information is valuable and can not be found in other literature that was associated with summative evaluations. This unique

feature should contribute to the use of the combination of the multiple-matched-sample and statistical controls.

B. Comparing Academic Achievement among Magnet Programs

The second question of this evaluation is: Among magnet programs, which were the most effective in improving students' academic performance in reading and mathematics after controlling for students' demographics and their initial abilities? Relative program effectiveness was based solely on the comparative magnitude of effect sizes that are scale invariant or metric free as indicated previously.

The effect size analysis that is summarized in Table 2 for all magnet programs indicates that the French Immersion (ES = 0.23) was the most effective among magnet programs in producing the larger effect size in MSA reading or mathematics performance. The rest of the magnet programs are not effective enough to be mentioned because their respective ES means were less than the cutoff value of 0.2.

Table 2
The Distribution of Effect Sizes for SY 2003 MSA Reading and Mathematics Test Scores for Each Elementary School Magnet Program (N of Replication = 200)

Content Area	Magnet Programs	Sample Size	Mean of ES	Standard Deviation	Minimum	Maximum	PR for the Mean ES
Reading	Academic Center	468	0.12	0.02	0.07	0.16	55
	Music & Technology	32	-0.16	0.16	-0.47	0.29	44
	Communication & Academic	269	-0.15	0.04	-0.24	-0.06	44
	Creative & Performing Arts	76	-0.17	0.09	-0.35	0.03	43
	French Immersion	55	0.23	0.09	-0.00	0.41	59
	Montessori	78	0.19	0.08	0.02	0.37	58
	Science, Math & Technology	405	-0.13	0.03	-0.20	-0.08	45
Math	Academic Center	468	-0.15	0.02	-0.21	-0.10	44
	Music & Technology	32	0.00	0.13	-0.30	0.28	50
	Communication & Academic	269	0.02	0.03	-0.08	0.11	51
	Creative & Performing Arts	76	0.10	0.09	-0.12	0.29	54
	French Immersion	55	0.32	0.09	0.08	0.59	63
	Montessori	78	-0.08	0.07	-0.22	0.12	46
	Science, Math & Technology	405	0.07	0.03	-0.02	0.15	53

% The PR stands for the percentile rank (PR) standing of the magnet -sample mean when it compares with the distribution of the matched-sample test scores.

IV. Summary and Conclusions

The summative evaluation of the magnet school programs employed a quasi-experimental design to investigate whether or not students enrolled in the magnet programs gained any achievement advantage over students who were not enrolled in these programs.

Researchers used *Zero-One Linear Programming* to draw multiple sets of matched samples from the non-magnet student population to serve as multiple control groups for this research effort. Each matched sample had three defining characteristics: (a) it was identical in size to the respective magnet group; (b) it was identical in demographic background (viz., gender, race, or poverty status itself as well as the combinations of those three variables) to its respective magnet groups; and (c) its average pretest score was very close (i.e., no statistically significant difference in most cases) to that of its respective magnet groups. Analysis of covariance (ANCOVA) procedure was used to control for the effects of any pre-existing difference in pretest scores between the two groups.

Whenever a matched sample was generated, the analysis of effect size was subsequently performed. Two hundred replicated matched samples were used in this evaluation. The mean of the effect size measures, across 200 replicated comparisons, was used to assess the program effect for the fourth graders enrolled in the magnet program treatment on their succeeding grade five reading and mathematics performance. This multiple-matched-sample control together with the ANCOVA control employed in this evaluation increases our confidence level in the findings. Moreover, the methodology used in this study is a particularly safe choice in evaluating any magnet program whose sample size is relatively small.

The results found in this study indicated that when students' demographics and initial abilities were taken into account, most magnet programs had minimal, if any, positive effect for raising reading and mathematics test scores of those magnet students, with the exception of the French Immersion program. The reasons why this program performed better than other ones is beyond the scope of this study. A further formative evaluation for this program may help us address this concern.

In drawing conclusions from the current study, it should be noted that causality may not be inferred from these analyses due to the lack of random assignment of students to the magnet and non-magnet student cohorts. Using random assignment of sample populations, two groups of students are more likely to be equated on all possible variables (Judd, Smith & Kidder, 1991). For the circumstance of the current policy for magnet student enrollment, the random assignment is impossible. The matched sample approach that has been employed in this study is closest to the random assignment approach, compared with other matched approaches that exist in the literature. In particular, the multiple-matched-sample approach prevents the possible bias caused by the selection of a matched sample. As the result of the combination of the multiple-matched-

sample and statistical controls, a more reliable and less biased estimate of program effectiveness is anticipated to be obtained

Finally, although the findings obtained from this study were based on sound evaluation design as well as an appropriate statistical analysis, magnet program effectiveness was assessed only by the performance of students in reading and mathematics. The issue of whether or not a specific magnet program (e.g., the French Immersion Magnet Program) has met its intended goals and objectives was beyond the scope of this evaluation study. Other magnet program effects (e.g., gaining knowledge of modern technology, having better self-esteem, etc.) cannot be addressed by this design. Accordingly, the findings obtained from this study only reflect the picture of academic performance of magnet programs, nothing else.

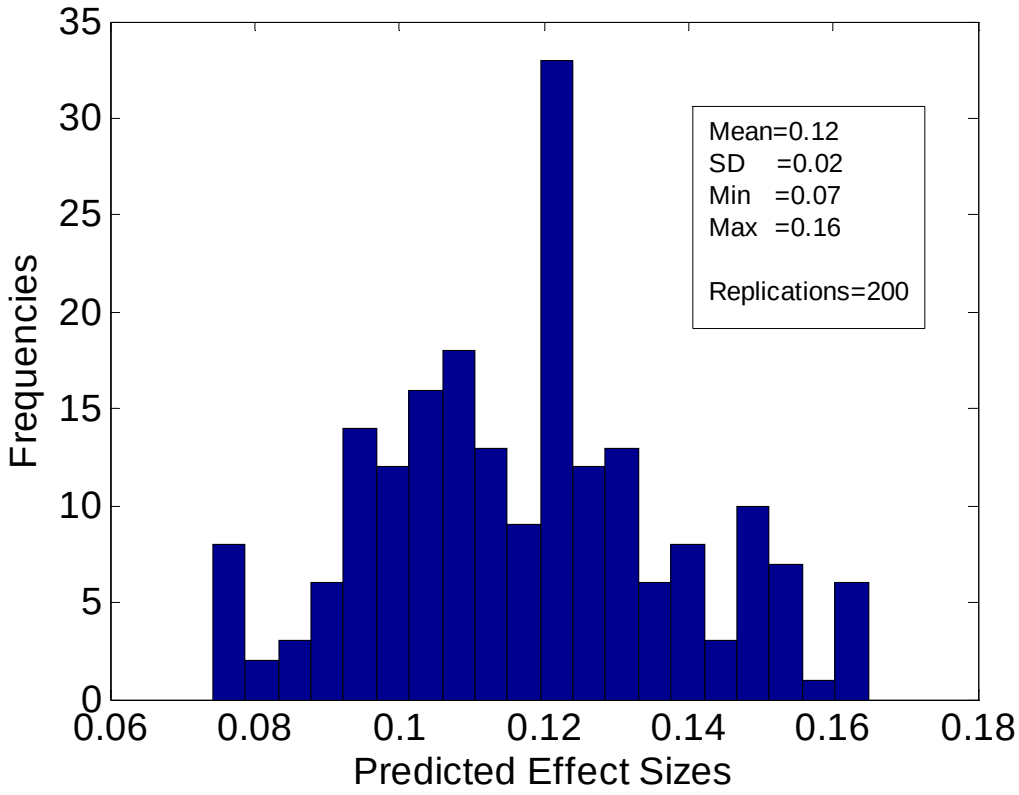

Figure 2. The Distribution of Effect Sizes of Reading for the Academic Program

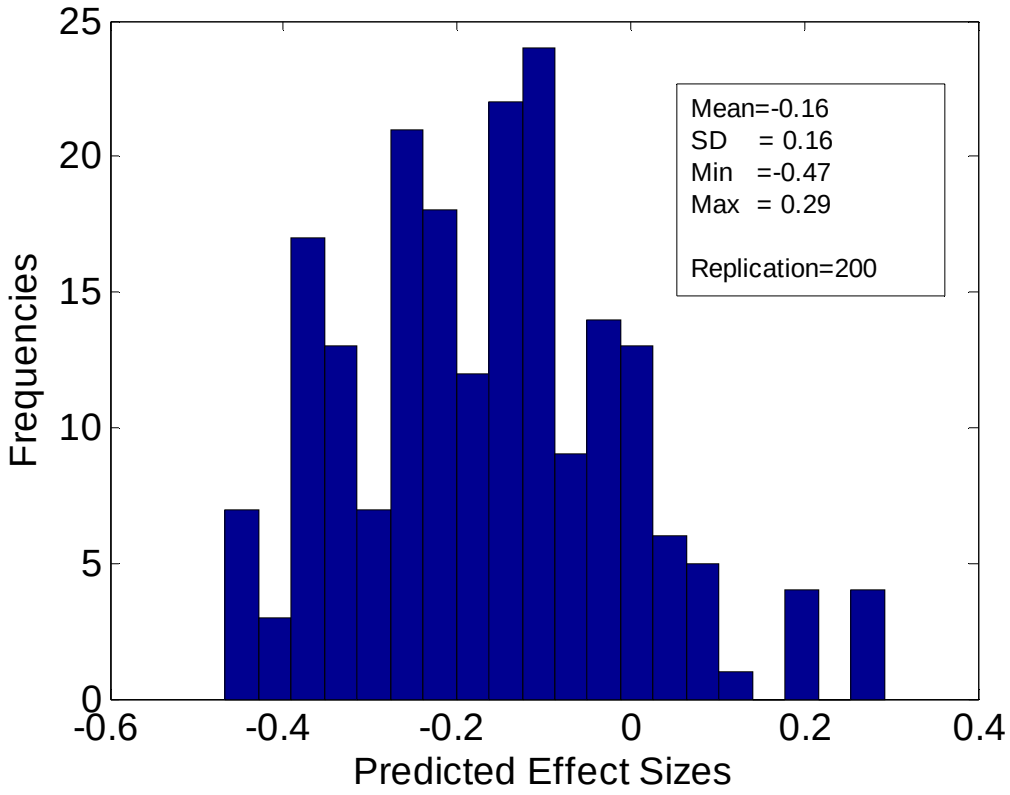

Figure 3. The Distribution of Effect Sizes of Reading for the Music Program

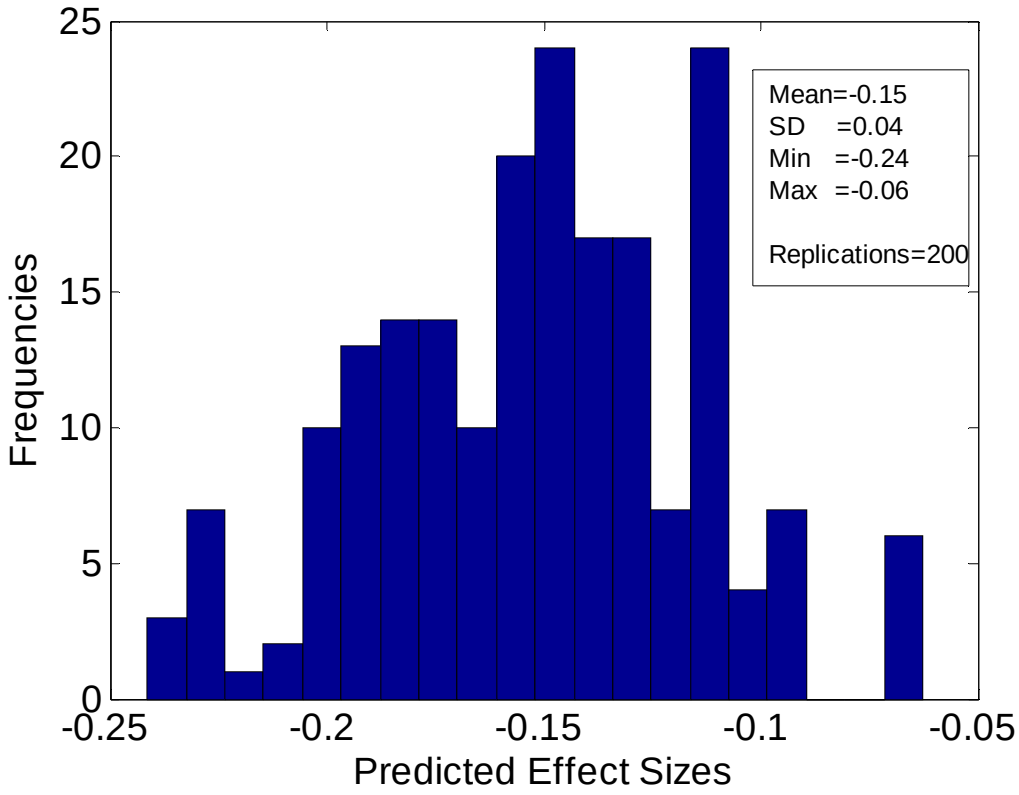

Figure 4. The Distribution of Effect Sizes of Reading for the Communication Program

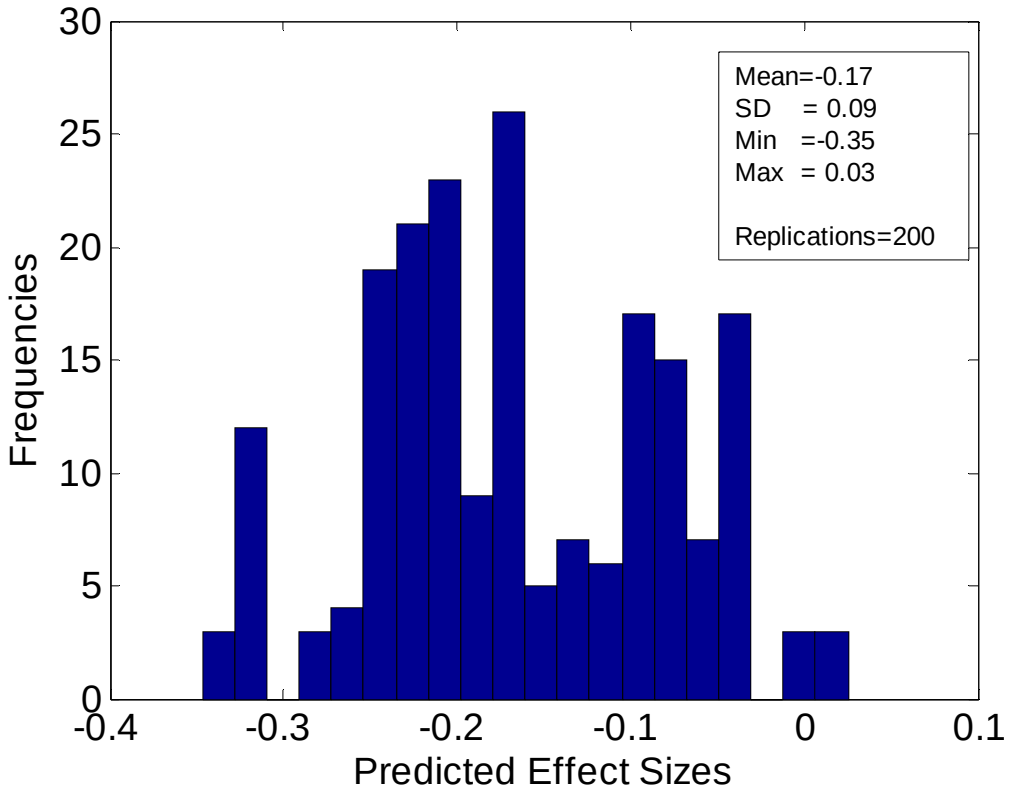

Figure 5. The Distribution of Effect Sizes of Reading for the Creative and Performing Arts Program

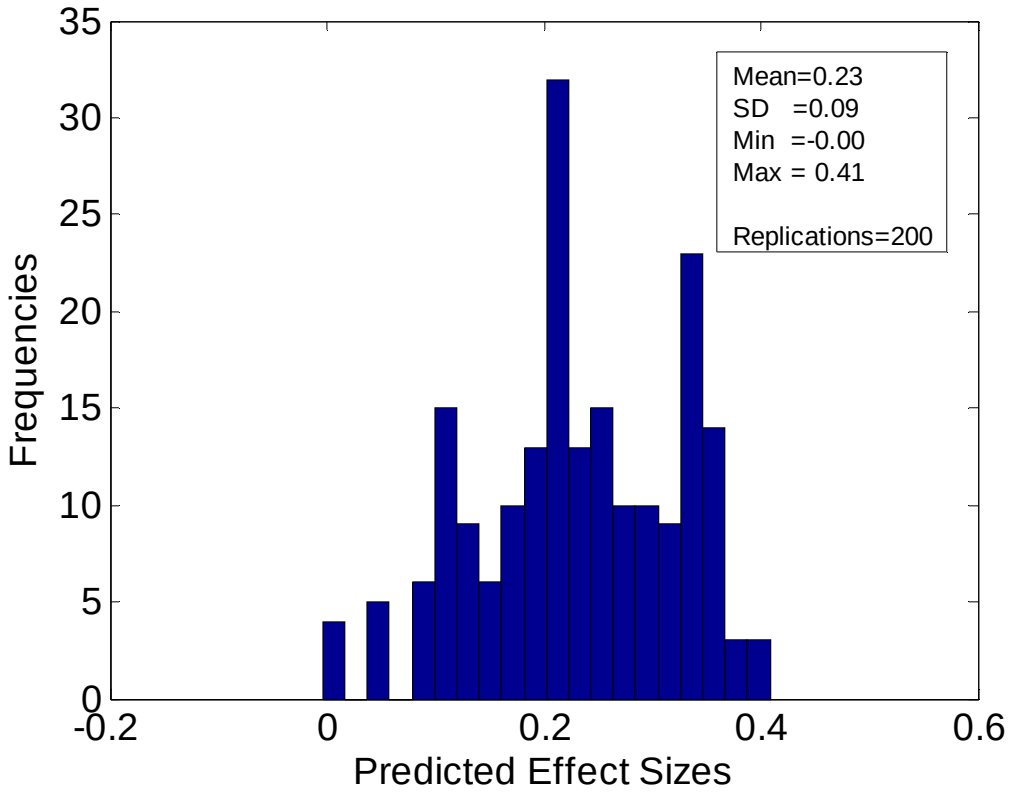

Figure 6. The Distribution of Effect Sizes of Reading for the French Immersion Program

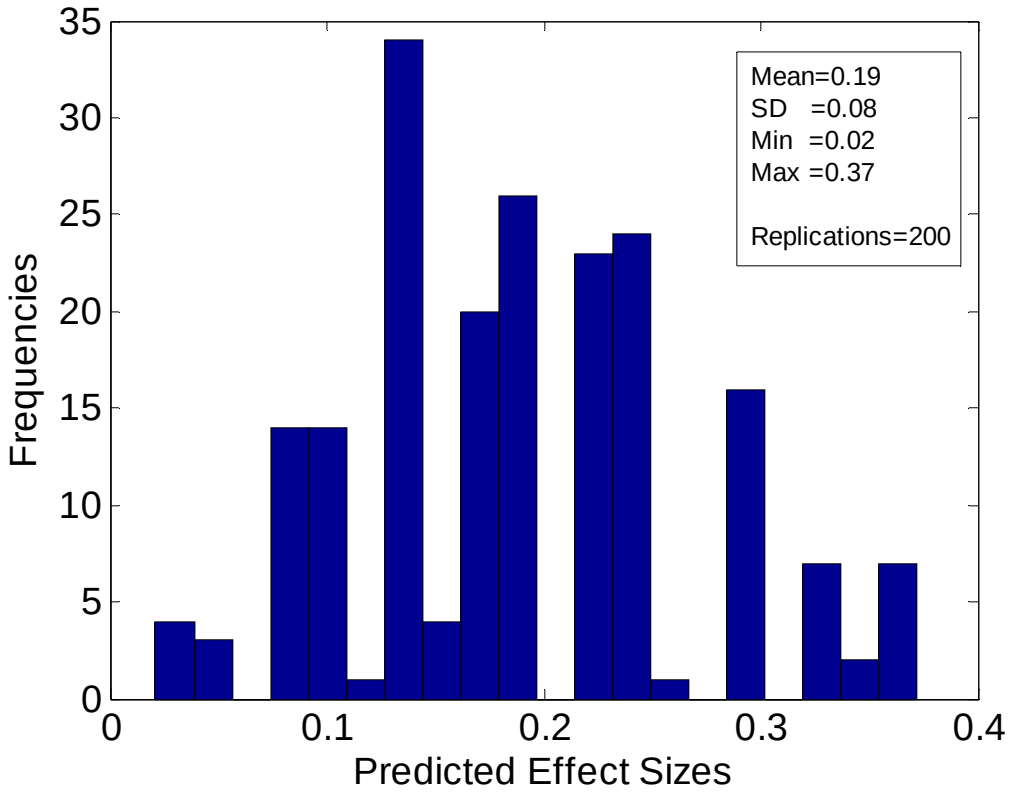

Figure 7. The Distribution of Effect Sizes of Reading for the Montessori Program

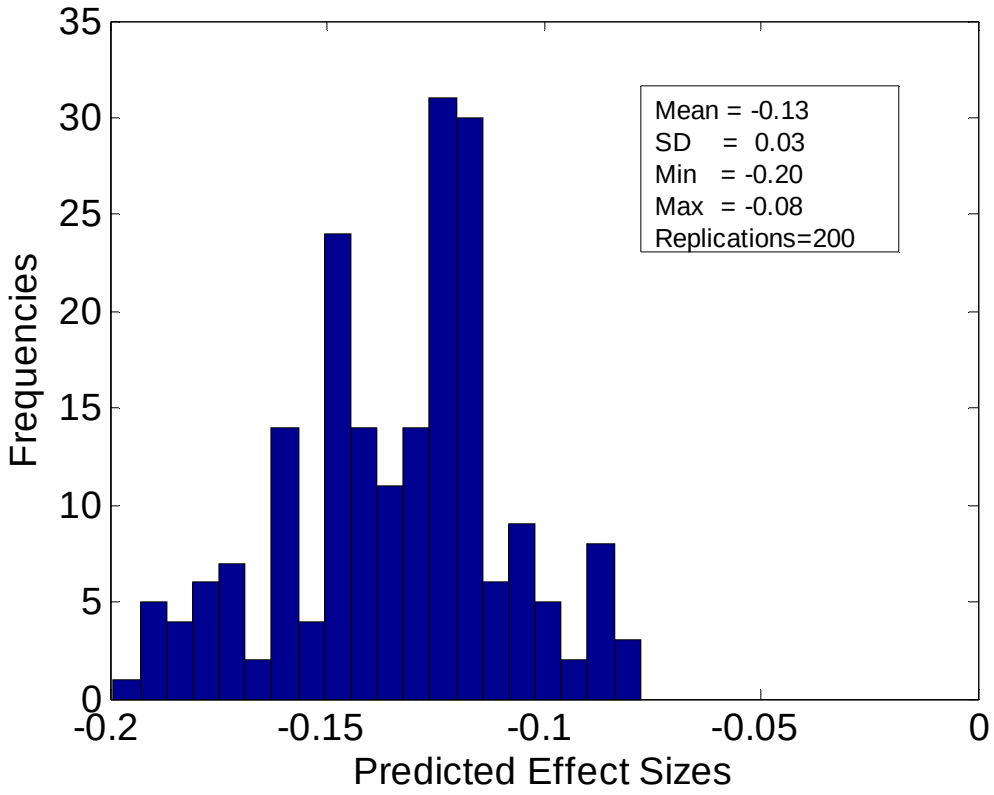

Figure 8. The Distribution of Effect Sizes of Reading for the Science & Technology Program

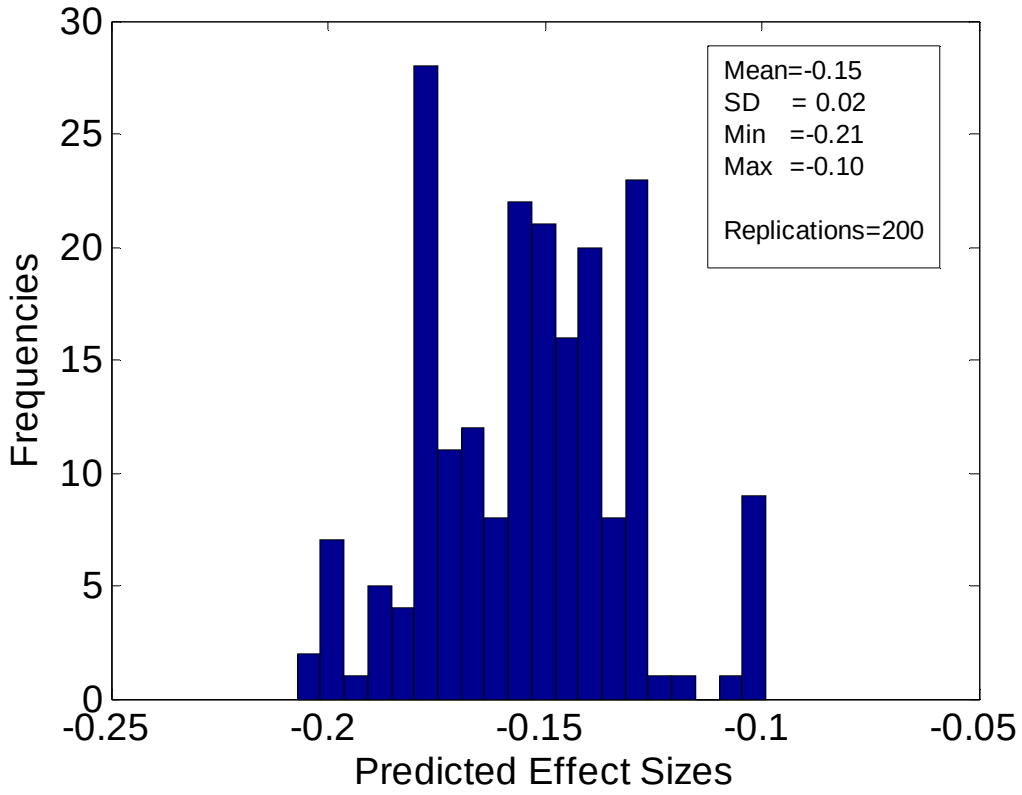

Figure 9. The Distribution of Effect Sizes of Mathematics for the Academic Program

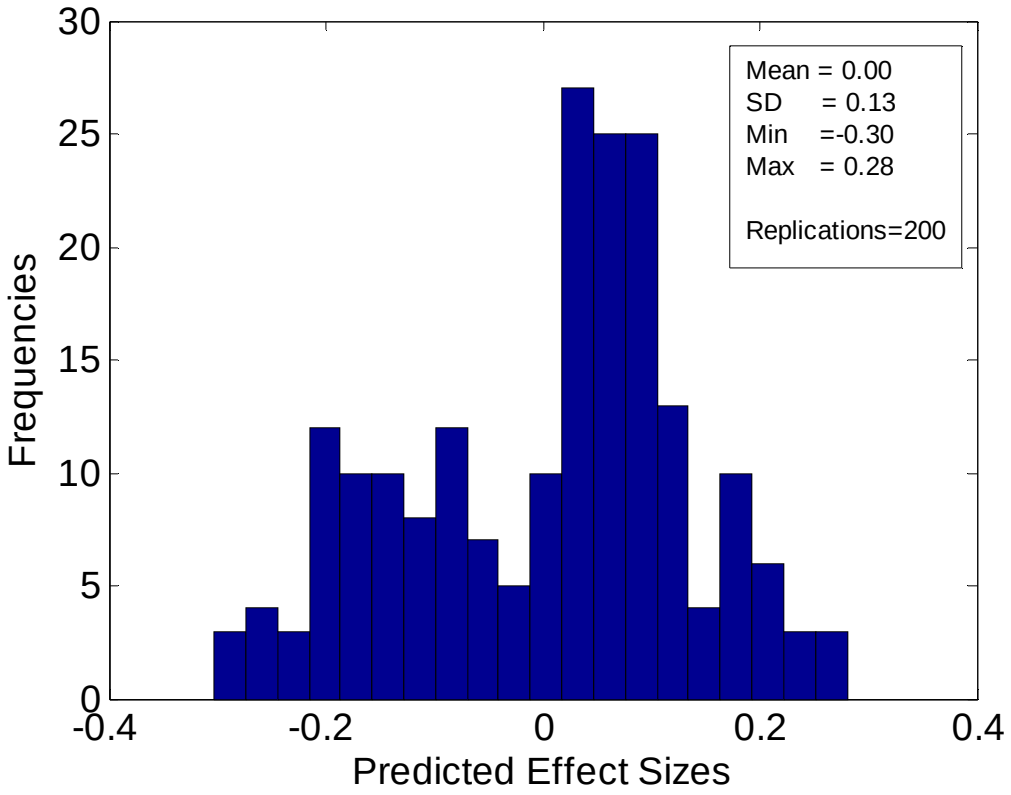

Figure 10. The Distribution of Effect Sizes of Mathematics for the Music Program

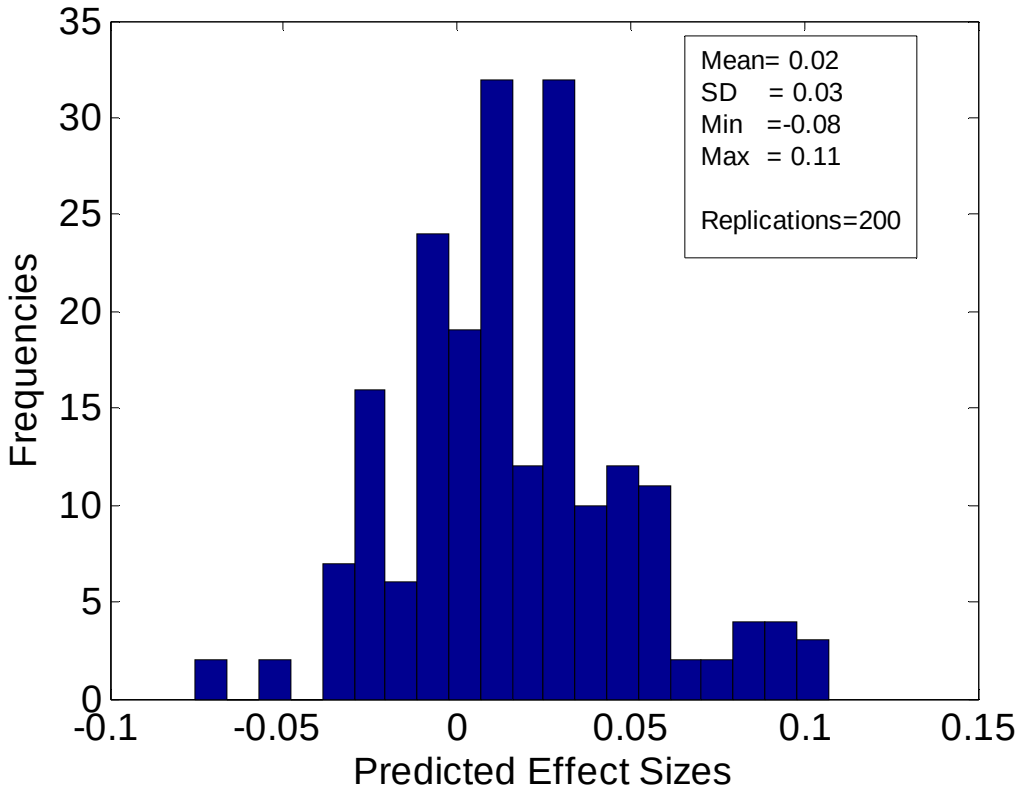

Figure 11. The Distribution of Effect Sizes of Mathematics for the Communication & Academic Program

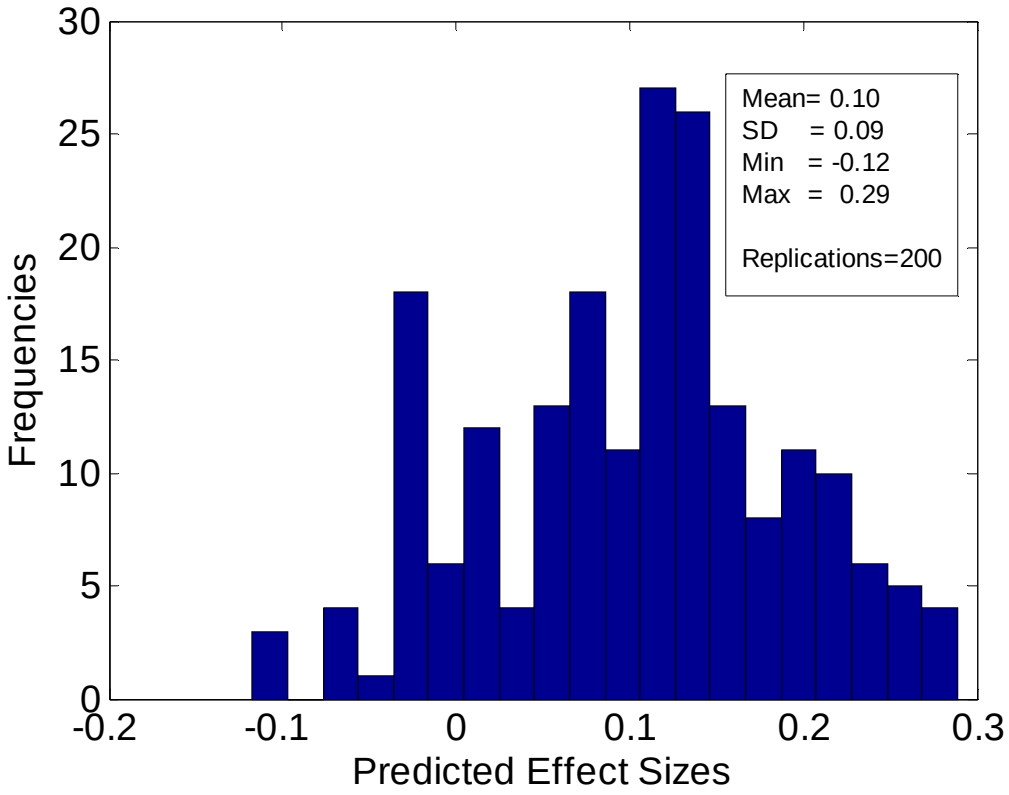

Figure 12. The Distribution of Effect Sizes of Mathematics for the Creative & Performing Arts Program

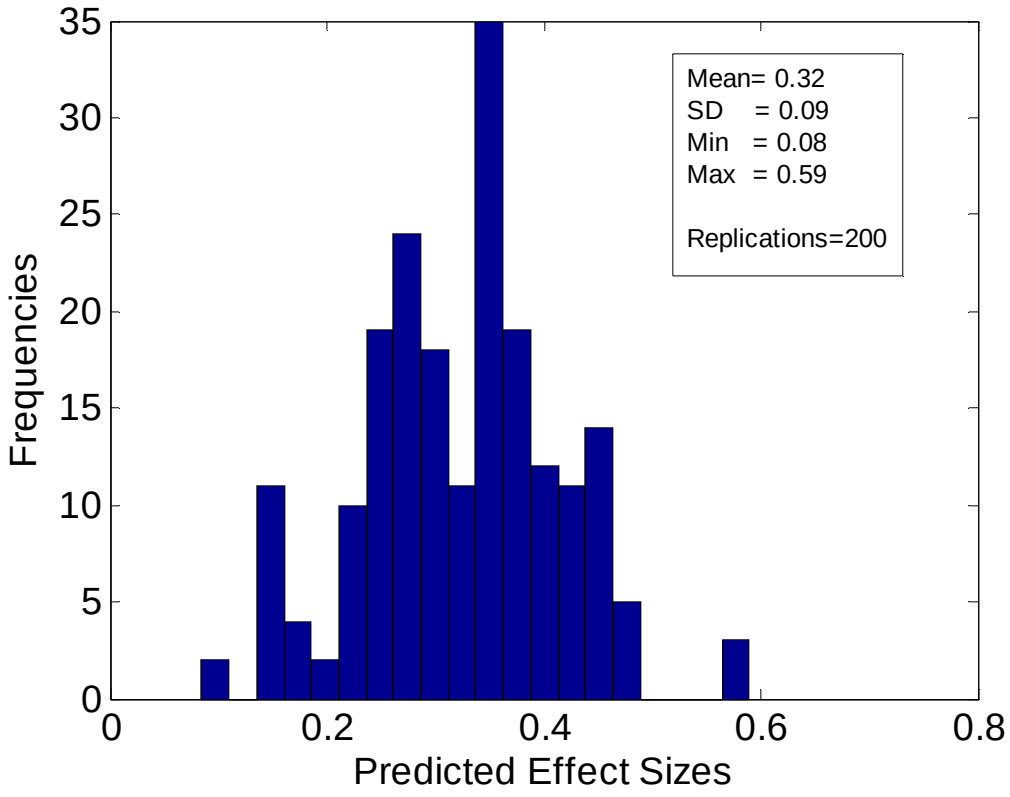

Figure 13. The Distribution of Effect Sizes of Mathematics for the French Immersion Program

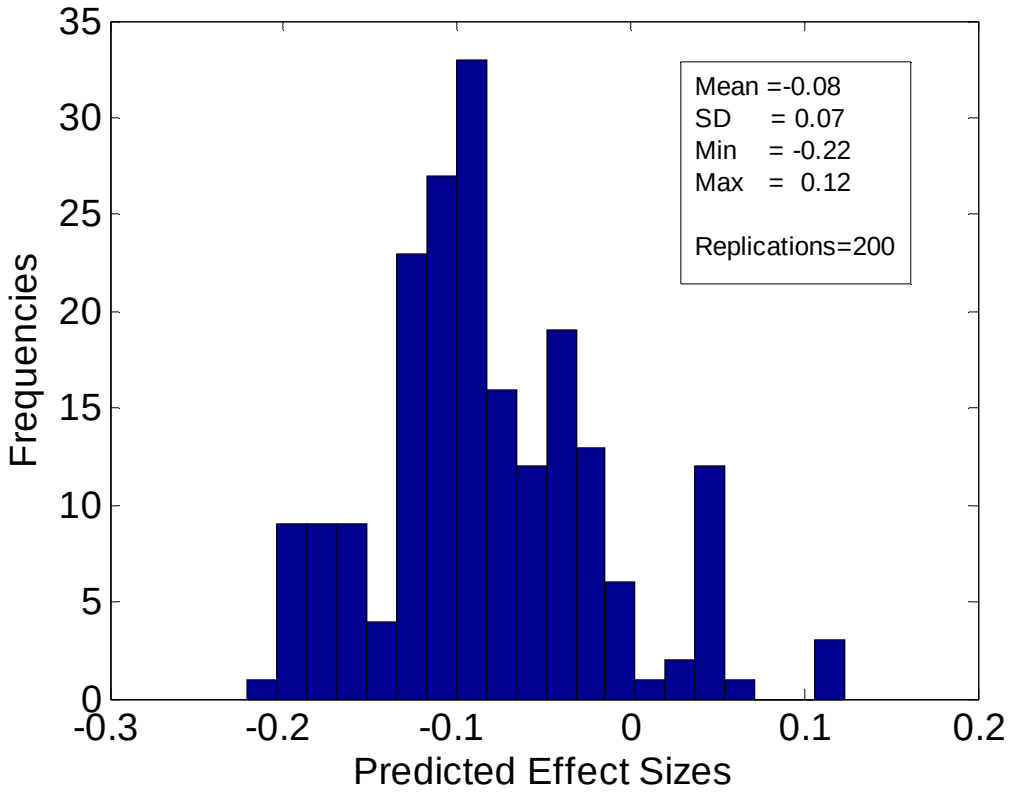

Figure 14. The Distribution of Effect Sizes of Mathematics for the Montessori Program

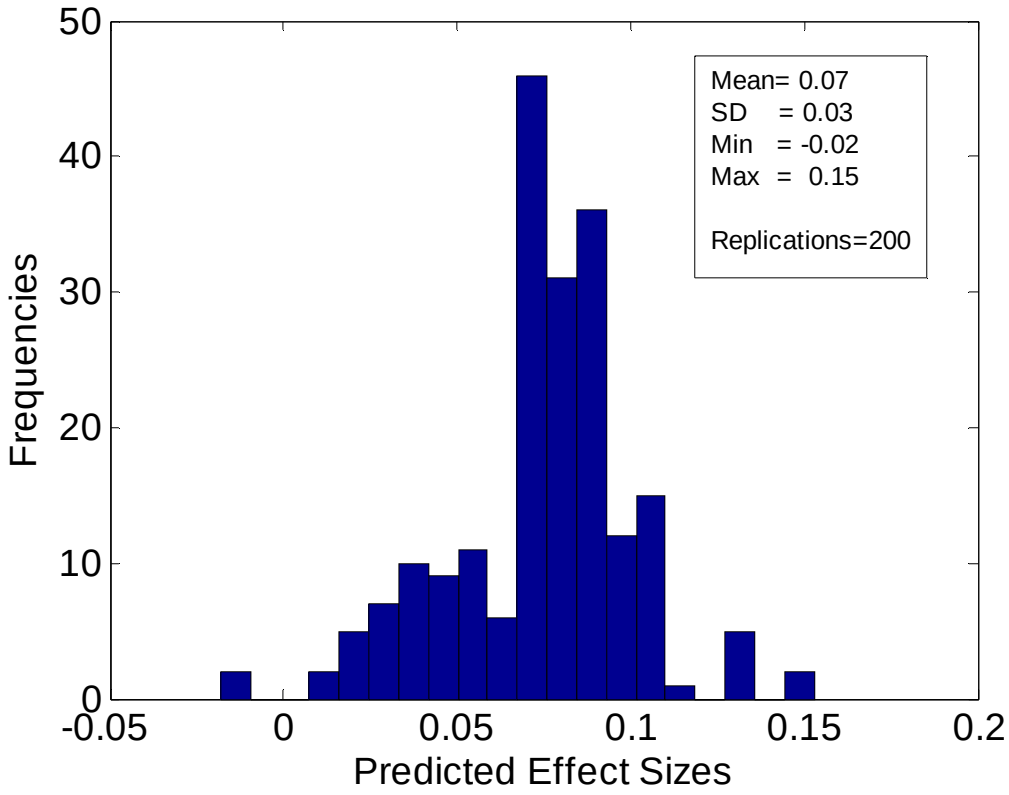

Figure 15. The Distribution of Effect Sizes of Mathematics for the Science, Math & Technology Program

References

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models: Applications and Data Analysis Methods. Newbury Park, CA: Sage Publications, Inc.

CTBS/McGraw-Hill (1997). Teacher's guide to TerraNova. Monterey, CA. McGraw-Hill Companies, Inc.

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

Isaac, S. & Michael, W. (1995). *Handbook in research and evaluation*, (3nd Ed.). EdITS / Educational and Industrial Testing Service, C.A.

Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). Research Methods in Social Relations. San Francisco: Holt, Rinehart, and Winston, Inc.

Li, Y. H. & Schafer, W. D. (2005a). Increasing the homogeneity of CAT's Item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests, *Journal of Educational Measurement*.

Li, Y. H. & Schafer, W. D. (2005b). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. Applied Psychological Measurement, 29, 1-23.

LINDO Systems, Inc.,(2003). *LINDO API: User's Manual*. LINDO Systems, Inc, Chicago, IL.

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 incorporate the propensity score. The American Statistician, 39, 33-38.

Rosenbaum, P. R. (1995). Quantiles in nonrandom samples and observational studies. Journal of the American Statistical Association, 90, 1424-1431.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. Journal of Educational and Behavioral Statistics, 29(1), 103-116.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton Mifflin Company, M.A.: Boston.

The MathWorks, Inc. (2003). MATLAB (Version 6.5): The language of technical computing [Computer program]. Natick MA: The MathWorks, Inc.

Theunissen, T. J. J. M. (1985). Binary programming and test design, *Psychometrika*, 50, 411-420.

Theunissen, T. J. J. M. (1986). Optimization algorithms in test design, *Applied Psychological Measurement*, 10, 381-389.

Thompson, B. (2002). "Statistical," "Practical," and "Clinical": How many kinds of significance do counselors need to consider?

van der Linden, W. J., & Boekkooi-Timminga, E. (1989).A maximum model for test design with practical constraints, *Psychometrika*, 54, 237-247.

Appendix A: Assumptions Underlying the Use of ANCOVA

The ANCOVA combines analysis of variance (ANOVA) and regression analysis. Several key assumptions underlying ANCOVA come from the two analyses and are: a) the samples are independent random samples from defined populations; b) the scores on the dependent variable are normally distributed in the population; c) the population variables in all cells of the factorial design are equal (homogeneity of variance); d) the relationship between dependent and covariate variables is linear; e) the regression lines (or slopes) for all groups (e.g., Reading Recovery students and matched sample) are assumed to be parallel (Hinkle, Wiersma, & Jurs, 1994). Failure to meet these assumptions will change the Type I error rate.

Generally, if the scores on the dependent variable are not normally distributed, or if the population variables in all cells of the factorial design are not equal, the Type I error rate is likely to change. When population samples are not normal (assumption (b) above), the Type I error rate can only be slightly changed by making the sample size larger than 20. If the population variance differs (assumption (c) above), the Type I error rate cannot be sizably changed by making the sample sizes equal. These findings are summarized by Hinkle, et. al., (1994) and can also be found in other literature (e.g., Cohen, 1988; Glass, 1984, Kirk, 1995, etc.).

Because most cognitive variables (e.g., test scores) are linearly related, and unless measurement instruments are faulty (e.g., ceiling effect), the linear relationship assumption (d) works well in most applications (Glass & Hopkins, 1984).

The parallel slopes (assumption e) might not be met in most data. The Monte Carlo study conducted by Glass, Peckham, and Sanders (1972) concluded that this violation has little effect on Type I error, although such a conclusion was not reached by other studies (e.g., Rogosa, 1980). When the slopes are parallel, we have more confidence to answer the question of whether there are differences between the two groups, by estimating the adjusted scores for the two groups at any given pretest score. The answer to this question should be the same for any given pretest score. However, when the slopes are not parallel, the magnitude of the adjusted-score difference between two groups depends upon which pretest score is selected. Because of the use of the *Zero-One Linear Programming* in the current study, we do not expect to find a sizeable difference in the pretest score between the matched sample and magnet program students. Small or no pretest difference between both groups will mitigate the problem of the possible violation of this assumption for each program evaluation.

A statistical test was used in this study to determine whether or not the difference in the ANCOVA-based adjusted scores between the two groups is significant. The Type I error for this statistical test could be changed because of the use of ANCOVA results; however, as explained above, a Type I error rate will not be dramatically changed even though some of the ANCOVA assumptions were not completely satisfied.

For this study, the effect size was primarily used to assess the magnitude of the program effect on student test performance. The index of the effect size is not a statistical test, so most assumptions (e.g., assumptions, a, b, and c) regarding the ANCOVA analysis are irrelevant to the procedure of computing the effect size values.

Appendix B: The Meaning of the Effect Size

This study sought to determine whether or not there was evidence that magnet program instruction had a positive impact on student achievement. ES was a method with which to judge the relative worth of programmatic treatment or non-treatment on the test performance of students from independent (or also in this case, matched) samples of students. This index can also be thought of as the Percentile Rank (PR) standing of the magnet program sample mean within the distribution of the matched-sample test scores (Cohen, 1988).

For example, if a particular magnet program treatment results in an effect size of (0.20), the area under the normal curve would be (0.58) or (0.5+(.08)). This would mean that the treatment effect would be expected to move a typical student in the treatment group from the 50^{th} percentile to the 58^{th} percentile of the control group. Using the rationale developed by Cohen (1988), a look-up table presented below was used to interpret the meaning of ES in terms of its PR standing in the matched sample (See Table A-1.)

Table A-1: Converted ES to Its Corresponding Percentile Rank (PR) Standing in the Matched Sample

Effect Size (ES)	Percentile Rank Standing
-0.5	31
-0.4	34
-0.3	38
-0.2	42
-0.1	46
0.0	50
0.1	54
0.2	58
0.3	62
0.4	66
0.5	69
0.6	73
0.7	76
0.8	79
0.9	82
1.0	84

Appendix C:

Utilizing the Zero-One Linear Programming Constraint to Draw Matched Samples from a Non-Treatment Population as Control Groups for a Quasi-Experimental Design

By

Yuan H. Li, Yu, N. Yang & Leroy J. Tompkins

Prince George's County Public Schools, Maryland

Shahpar Modarresi

Montgomery County Public Schools, Maryland

Paper was presented at the annual meeting of the American Educational Research Association, Montreal, Canada, April, 2005.

Utilizing the Zero-One Linear Programming Constraint to Draw Matched Samples from a Non-Treatment Population as Control Groups for a Quasi-Experimental Design

Abstract: The statistical technique, Z*ero-One Linear Programming*, that has successfully been used to create multiple tests with similar characteristics (e.g., item difficulties, test information and test specifications) in the area of educational measurement, was deemed to be a suitable method for creating matched samples to be used as control groups in the quasi-experimental design of *non-randomized comparison group pretest-posttest*. Compared to the existing propensity-score matching method, this method does not require any statistical models and assumptions and can handle the covariate of the pretest score more appropriately.

If the measurement error of the pretest-score mean of the treatment group is ignored, this method will generate a *unique* matched sample once the criteria for attempting to create two similar groups are determined. Otherwise, multiple similar matched samples can be generated and the performance of the treatment group can be compared with each of the multiple matched samples using an appropriate statistical analysis. Afterwards, the mean of the effect size measure, taking the average of the effect size across replicated comparisons, can then be used to assess the efficacy of any program. This enhances our confidence level to decide whether a program is effective or not, compared to the finding resulting from a single comparison.

A description of Z*ero-One Linear Programming* and its application to create a matched sample or multiple matched samples is introduced in this paper.

Key Words: Linear Programming, Matched Sample, Quasi-Experimental Design
Optimization, Experimental Design, Program Evaluation

I Introduction

A. Background of Quasi-experimental Design

In an experimental design, random assignment is an ideal sampling method to create experimental and control groups when a group of subjects is available. The subjects of the experimental group will receive a treatment; whereas, no specific treatment will be given to the subjects of the control group. The procedure of random assignment becomes a powerful technique for controlling all known and "unknown" extraneous variables because it makes both groups very similar at the beginning of an experiment, especially in cases where the sample size is large. Unfortunately, this method has often encountered implementation obstacles in the evaluation of educational programs because student enrollment in a specific program is not random, and as such, cannot be completely manipulated as can be done with random assignment in most instances. Accordingly, the quasi-experimental design, defined as an experiment without randomized assignment but involving the manipulation of independent variables (Isaac & Michael, 1995; Shadish, Cook & Campbell, 2002), becomes one of the alternatives used to determine a program effect.

The *non-randomized comparison group pretest-posttest design* (illustrated in Figure 1, refer to Shadish, Cook & Campbell [2002], p. 136) is the most appropriate evaluation design in assessing the efficacy of any program among the quasi-experimental designs. For this design, random assignment is not conducted and subjects in both the quasi-experimental and the control groups will take both the pretest and the posttest. Like a true experimental design, the subjects in the control group will not receive any specific treatment, but their counterparts in the quasi-experimental group(s) will receive program treatment(s). Here, the number of groups under the quasi-experimental label could be single (e.g., only one program) or multiple (e.g., several programs to be evaluated simultaneously).

Without random assignment in the design introduced above, the impact of undetected nuisance variables on the outcome variable might not be ruled out. In order to better assess the efficacy of a program, a purely statistical modeling (e.g., analysis of covariance, ANCOVA, Kirk, [1995]; hierarchical linear modeling, HLM, Bryk & Raudenbush, [1992]) may be used to tackle this issue. Without using the matched sample procedure, the statistical modeling is

primarily used to account for student's characteristic differences (e.g., sex, race, pretest scores, etc.) or school context differences (e.g., percent of minority students, percent of poverty students, etc.) between treated and non-treatment groups. However, as pointed out by Rubin, Stuart and Zanutto (2004), comparing results obtained from treated (e.g., magnet programs) and whole control groups (e.g., non-magnet population) with very different distributions of background covariates will heavily rely on untestable modeling (e.g., ANCOVA) assumptions and extreme extrapolation. As such, reliable causal inferences may not be drawn. For example, Rubin et. al. (2004) further illustrated that the values of "percent minority" and "percent in poverty" may differ widely at some schools, this situation will cause the estimated school effects that have been adjusted for such covariates using models be extremely sensitive to these statistical modeling assumptions (e.g., parallel slopes). If the assumptions are seriously violated the distributions of background variables among subgroups are different to some extent, the estimated program effect, as a result of using extreme extrapolation, will be seriously misleading.

	Pretest	Treatment		Posttest	
Quasi-Experimental Group(s)	O_1	=>	X	=>	O_2
Control Group	O_1	=>	C	=>	O_2

where,
O_1 – Pretests
X – Treatment(s)
C – No Treatment
O_2 – Posttests

Figure 1: The Non-Randomized Comparison Group Pretest-Posttest Design

B. Issues Associated with Creating a Matched Sample

The goal of obtaining a *Matched Sample as a Control Group* is to create the conditions similar to a randomized experiment as closely as possible. The treated and control groups are matched without using the observed outcome variable (or posttest), thus preventing us from "intentionally" manipulating a matched sample to obtain a desired result and also protecting from such claims by researchers. The ability of a matched sample procedure to reveal the extent to which treated and matched groups have similar types of students in similar educational settings is "an important diagnostic tool to identify whether the data can support [possible] causal comparisons between these two groups" (Rubin et. al., 2004).

Instead of totally utilizing statistical modeling in assessing program effects, it would be preferable to use an appropriate matched sample to be compared with the treatment group before any statistical modeling is performed. However, creating an appropriate matched sample for each program is a major challenge. Diverse methods can be used to accomplish this objective (for literature review, see Shadish et. al., 2002). Among them, using the propensity score (Rosenbaum & Rubin, 1983, 1984, 1985; Rosenbaum, 1995) as a criterion for selecting students as members of a matched sample is one of the promising approaches to address this issue. A propensity score is an estimated probability of a given individual belonging to a treatment group given the observed background characteristics (or covariates) of that individual. This propensity score reduces the entire collection of background characteristics to a single composite index value so that a matched sample will be selected using this single index, instead of directly matching multiple background variables.

Nevertheless, the value of the propensity score is dependent on the selections of statistical models (e.g., whether or not including the interaction, and/or nonlinear terms on the logistic regression models). Also, if the assumptions made for the statistical model (e.g., logistic regression) are not met, and/or if the sample size used for the model is not large enough, using those propensity scores as a criterion for selecting a matched sample might not be as meaningful as researchers anticipated. Furthermore, the weighting for each covariate, that is then used for computing the propensity score, depends totally on the degree of each covariate's relation to the treatment assignment (received or not received treatment). This procedure is not appropriate for the *non-randomized comparison group pretest-posttest design*, in which the pretest score is usually highly correlated with the outcome measure rather than the treatment assignment. This scenario will cause the pretest-score covariate to be less important than it should be when the propensity-score is used to select a matched sample.

Due to some limitations of the propensity score method, this paper is intended to introduce another procedure to create matched samples for this *non-randomized comparison group pretest-posttest design*. Several studies (e.g., Li and Schafer [2005a], Li and Schafer [2005b], Theunissen [1985 and 1986], and van der Linden and Boekkooi-Timminga [1989])

have successfully utilized the Z*ero-One Linear Programming* technique in the area of educational measurement for creating multiple tests with similar characteristics (e.g., item difficulties, test information, and test specifications). Using similar logics, this technique is also appealing when it is used to create a matched sample for this design. The matched sample method introduced in this paper has potential to be applied to other quasi-experimental designs with some modifications. A description of Z*ero-One Linear Programming* is presented below, before detailed steps of using this technique in creating a matched sample as a control group are introduced.

II. Introduction of Zero-One Linear Programming

The techniques of optimization help us seek the solution that provides the best result (e.g., attaining the highest profits while making the most efficient use of our resources including money, time, machinery, staff, inventory, and more). Such problems are often classified as linear or nonlinear, depending on the nature of the relationship of the variables involved in the problem (LINDO Systems, Inc., 2003).

A. Zero-One Linear Programming

Linear programming is designed to seek the maximum (or minimum) value for a linear function such as in Equation 1, while the required constraints, formalized in Equations 2 and 3, are imposed.

$$\text{Minimize } \sum_{i=1}^{n} \left[\text{ABS}(P_i - M)\right] x_i \qquad (1)$$

Such that

$$\begin{aligned} &A_{11}x_1 + A_{12}x_2 + \ldots + A_{1n}x_n \square b_1 \\ &A_{21}x_1 + A_{22}x_2 + \ldots + A_{2n}x_n \square b_2 \\ &\quad\vdots \qquad\quad \ldots \qquad\quad \vdots \end{aligned} \qquad (2)$$

$A_{m1}x_1 + A_{m2}x_2 + \ldots + A_{mn}x_n \square b_m$

$x_i \in \{1,0\}$ (3)

where

ABS is the absolute function,

P_i is the pretest scores for all members without receiving a specific treatment,

M is the pretest score mean of the quasi-experimental group,

i is the member index for all members without receiving a specific treatment (i=1,…,n),

A_{mxn} is the coefficient, and b_m is the right-hand side value for the m^{th} constraint (to be delineated in the example below),

$\square$ is the relationship function, which could be ≤, =, or ≥. The equal symbol of '=' is used here.

More specifically, in Equation 1, the members in the population are indexed by i=1,…, n and the values in the variable x_i are parameters that will be estimated. For *zero-one linear programming*, the x values are constrained to be either one or zero as indicated in Equation 3 to identify whether the members are selected or not for the matched group.

Equation 2 introduced above can be presented by a matrix expression—Equation 4 (shown below) in which the vector of $\underline{x}$ will be resolved by not only maximizing (or minimizing) the linear function of Equation 1, but also imposing the constraint of x values of either one or zero. The matrix **A** and the vector $\underline{b}$ in Equation 4 are created from A_{mxn} and b_m coefficients, respectively. The way of preparing both matrix **A** and the vector $\underline{b}$ depends on the nature of the problem we attempt to resolve. It is noted that the following descriptions in illustrating how to prepare both matrix **A** and the vector $\underline{b}$ for the solution of zero-*one linear programming* only fit the problem presented in this paper. Readers might refer to other references (e.g., Theunissen, 1985, 1986) for better understanding this issue.

$\mathbf{A} \cdot \underline{x} = \underline{b}$ (4)

It is noted that if multiple pretest scores are available, a composite score obtained from those pretests is more appropriate to be entered into Equation 1. The choice of types of composite score can be dependent on the nature of those pretest scores themselves.

B. Example of Using Zero-One Linear Programming

In the present example, suppose two key student demographic variables (e.g., gender and poverty) are considered for matching. Under this circumstance, there are four types of students as shown in Table 1 – male/poverty, male/non-poverty, female/poverty, and female/non-poverty. It is further assumed that there are 10 and 20 students in the magnet and non-magnet (or comprehensive) programs, respectively, and the frequencies of each type of student are also shown in Table 1. Ten non-magnet students will be drawn from across the four student subgroups to correspond with the 10 magnet students. The number of non-magnet students drawn from each subcategory will be identical to the number of magnet students in the respective subcategory. At the same time, the average pretest scores between two groups are expected to be as close as possible.

Table 1. An Example of Data Regarding the Frequency of Four Types of Students

Type of Students	**Magnet Students Frequency**	**Non-Magnet Students Frequency**
1. male/poverty	2	4
2. male/non-poverty	3	6
3. female/poverty	1	2
4. female/non-poverty	4	8
Subtotal	**10**	**20**

Under the sampling scenario cited above, the **A** matrix and $\underline{b}$ vector in Equation 2 or 4 will be created as shown below before seeking the "Zero-One" solution of the vector parameter $\underline{x}$.

$$\mathbf{A}=\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\underline{b} = \begin{bmatrix} 2 \\ 3 \\ 1 \\ 4 \end{bmatrix}$$

Regarding the **A** matrix**,** the number of columns in the **A** matrix should equal the number of non-magnet students in the pool. In addition, each row in the **A** matrix together with the corresponding row in the $\underline{b}$ vector expresses a single constraint. The first constraint is expressed in the first row in the **A** matrix together with the first row in the $\underline{b}$ vector. The four series of "1" connotes that the first four of the 20 students are Type 1 students, and the rest of the sixteen series of "0" connotes that they are not Type 1 students. Further, the condition of two members in Type 1 students to be picked as part of a matched sample is specified as "2" in the first row in the $\underline{b}$ vector.

The second constraint is expressed in the second row in the **A** matrix together with the second row in the $\underline{b}$ vector. The six series of "1" connotes that they are Type 2 students and the rest of the fourteen of "0" connotes that they are not Type 2 students. Further, the condition of three members in Type 2 students to be picked as part of a matched sample is specified as "3" in the second row in $\underline{b}$ vector. Using the same logic, the third and fourth constraints are specified in the **A** matrix and the $\underline{b}$ vector.

After the **A** matrix and the $\underline{b}$ vector are set up and both are then inserted into Equation 4 ($\mathbf{A} \cdot \underline{x} = \underline{b}$), a specific mathematical function is formed and shown in Equation 5. The solution of *Zero/One* values of x_i parameters in this function will be found on the condition that the targeted function of Equation 1 is minimized in this case.

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{15} \\ x_{16} \\ x_{17} \\ x_{18} \\ x_{19} \\ x_{20} \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 1 \\ 4 \end{bmatrix} \quad (5)$$

As already indicated, a statistical technique, *Zero-One Linear Programming*, is used to seek the solution of the vector of $\underline{x}$. Afterwards, a matched sample will be created using the information of the x *zero/one* values. This matched sample will have the same demographic characteristics (e.g., poverty status, gender) as the magnet student population because the constraints formalized in Equations 2 and 3 are imposed. Furthermore, its average pretest score is close to the average pretest score of the magnet population because the linear function in Equation 1 is minimized.

It is important to note that "the same demographic characteristics" in those selected matching variables (e.g., poverty and gender variables) not only means that the number of students in each selected variable itself is identical between two groups, but also means that the

distribution of students in the combination (e.g., four types of students cited above) of those selected matching variables (e.g., poverty and gender) is identical. The latter feature has been demonstrated in the above example, but is too complicated to be done using previously existing matching procedures.

For the example introduced above, only the category variables (e.g., gender, race) were selected to create a series of constraints formulized in Equation 2. Other continuous variables (e.g., age) are also suitable to be included (refer to Theunissen, 1985, 1986).

III. A First Study Using the *Zero-One Linear Programming* Approach to Create a Matched Sample

An investigation of the effects of magnet school programs on the reading and mathematics performance of students in a school system was conducted by Yang, Li, Modarresi and Tompkins (2003). The major objective of this summative evaluation was to compare the academic performance of students from each of the magnet programs, to the performance of a matched sample of their non-magnet peers in the reading and mathematics content areas. The *non-randomized comparison group pretest-posttest design* has been used for this evaluation. Refer to Figure 1, the quasi-experimental group was a group of magnet program students, and the control group was a matched sample that was drawn from the population of non-magnet students using the *Zero-One Linear Programming* technique introduced above. The magnet program group received the magnet program treatment, while the non-magnet group did not.

The Comprehensive Test of Basic Skills (CTBS, CTB/McGraw-Hill, [1997]) reading and mathematics tests administered in 2001 were used as pretests for both groups of students. The 2003 Maryland School Assessment (MSA) reading and mathematics assessments were used as posttests. Measuring the magnet program treatment effect was of primary interest in the study cited above. The posttest score difference between two groups might be used for this purpose; however, the pre-existing difference between the two groups (e.g., initial abilities and

demographic differences) was not accounted for by only observing the simple posttest score difference. Most researchers often suggest that, if possible, both statistical and matching controls should be simultaneously employed in order to better adjust for those pre-existing differences at the beginning of the experiment. Such a principle was fully applied on that study, in which a matched sample was created for each magnet program before the ANCOVA was used to adjust for the small difference in the pretest score between the magnet and respective matched groups. The process of creating a matched sample (e.g., for the Academic Center Magnet Program) is described below.

A. Tabulate the frequencies of various types of students

As seen in Table 2, there were 468 and 6,435 fifth graders in the Academic Center Magnet Program and comprehensive (or Non-Magnet) programs, respectively. Students were grouped by combinations of race, gender, and poverty status, i.e., 20 types of students were classified and listed in the first column of Table 2. The frequencies of those 20 types of students are also shown in Table 2 for both Magnet and Non-Magnet groups.

Table 2.
Frequencies for the Academic Center Magnet Program and Non-Magnet Students

Types of Students	Magnet Program Frequency	Non-Magnet Students Frequency
1. American Indian, male, non-poverty	1	5
2. American Indian, male, poverty	0	7
3. American Indian, female, non-poverty	3	9
4. American Indian, female, poverty	0	8
5. Asian, male, non-poverty	3	53
6. Asian, male, poverty	3	34
7. Asian, female, non-poverty	7	55
8. Asian, female, poverty	2	32
9. African American, male, non-poverty	92	1,098
10. African American, male, poverty	80	1,399
11. African American, female, non-poverty	100	1,121
12. African American, female, poverty	103	1,548
13. White, male, non-poverty	19	215
14. White, male, poverty	0	46
15. White, female, non-poverty	19	219
16. White, female, poverty	4	44
17. Hispanic, male, non-poverty	9	35

18. Hispanic, male, poverty	15	210
19. Hispanic, female, non-poverty	2	46
20. Hispanic, female, poverty	6	251
Total	**468**	**6,435**

B. Choose the test score to be minimized

Since a matched sample is to be drawn and then applied to either reading or mathematics performance evaluation for this magnet program, the average pretest score of both CTBS reading and mathematics T scores was used to be minimized in the context of *Zero-One Linear Programming*. The T score equals (50 + 10 times z), where z is the standard score of the reading or mathematics "scale" score. Averaging T scores in reading and mathematics, instead of averaging their scale scores, is done to ensure that the weighting in both content areas is equal when both scores were added up together and then were averaged.

C. Utilize the Zero-One Linear Programming

Once each student's average pretest score and the distribution of various types of students for the Academic Center are available, the linear function (presented in Equation 1), matrix **A** and the vector $\underline{b}$ can be created. The *zero-one linear programming* then used them to seek the solution of the $\underline{x}$ vector indicated in Equation 1. The $\underline{x}$ vector was then used to identify which students were chosen to be part of the matched sample from among 6,435 non-magnet students. The frequency distributions and average pretest scores for the Academic Center and its matched sample are provided in Table 3, where it shows that the distributions of various types of students between the magnet program and matched groups are identical. Furthermore, their average pretest scores are almost the same. No statistically significant difference was found in the average pretest scores between the two groups.

Table 3.
Final Results: Frequency Distributions and Average Pretest Scores for the Academic Center and Its Matched Sample

Types of Students	Magnet Program Frequency	Matched Sample Frequency
1. American Indian, male, non-poverty	1	1
2. American Indian, male, poverty	0	0
3. American Indian, female, non-poverty	3	3
4. American Indian, female, poverty	0	0
5. Asian, male, non-poverty	3	3
6. Asian, male, poverty	3	3
7. Asian, female, non-poverty	7	7
8. Asian, female, poverty	2	2
9. African American, male, non-poverty	92	92
10. African American, male, poverty	80	80
11. African American, female, non-poverty	100	100
12. African American, female, poverty	103	103
13. White, male, non-poverty	19	19
14. White, male, poverty	0	0
15. White, female, non-poverty	19	19
16. White, female, poverty	4	4
17. Hispanic, male, non-poverty	9	9
18. Hispanic, male, poverty	15	15
19. Hispanic, female, non-poverty	2	2
20. Hispanic, female, poverty	6	6
Total N	**468**	**468**
Average Pretest Score	**50.53**	**50.52**
Difference of Pretest Score	**0.01***	

* P = .986

IV. Individually-based Matching

The matched sample generated by the above steps did not specifically identify the respective matched member given a treatment group (e.g., the magnet-program) member. As already indicated, this group-based matching procedure has been incorporated into the

magnet program study (Yang, Li, Modarresi and Tompkins, 2005). In some instances, a matched sample generated by an individual matching procedure is preferred. For example, when the program effect is analyzed by the whole group and is then required to be analyzed by the disaggregated subgroups, an individual matching procedure makes it possible that each subgroup has its own matched sample to be compared with.

This section further discusses the steps on how to modify current matching procedure to serve the purpose mentioned above. There are several possible solutions. The algorithm described below is one of the promising procedures.

(1). Start with the first individual member from the magnet group.
(2). Utilize the current group matching procedure to draw a matched sample.
(3). Find a member from the matched sample generated by Step 2 with the following conditions, a) Have the same type of member as this individual member indicated in Step 1, b) Have the closest pretest score to this individual member indicated in Step 1. Once this member is found, he/she will be the respective member of this individual member indicated in Step 1.
(4). Replace any members (note: those drawn by Step 2) that have not been previously selected by Step 3 in the non-magnet population pool.
(5). Add an additional constraint to the constraints that have been imposed in the *zero/one* linear model to ensure that the member being recently selected in Step 3 "must" be included in the next new-drawn matched sample. For matrix expression in the above example, if the second member from the non-magnet population is selected to be matched with the first member of the magnet group, the **A** matrix and $\underline{b}$ vector should be updated in the following way:

$$\mathbf{A}=\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\underline{b} = \begin{bmatrix} 2 \\ 3 \\ 1 \\ 4 \\ 1 \end{bmatrix}$$

The last row in **A** matrix combined with the last row in $\underline{b}$ vector indicates that the second member is certain to be one member in the next new-drawn matched sample.

(6). Repeat the steps 1-5 until all members in the magnet group have found their own respective members from the non-magnet population.

The above individually-based matching procedure begins with drawing a matched sample that meets all the desirable constraints and has a minimum pretest-score difference between both matched and respective treatment groups. A member who meets the criteria indicated in Step 3 is then selected from this matched sample, instead of directly from the full non-treatment population. Each of the next serial matched samples includes all previously selected members and consequently the last matched sample is an actual matched sample which always meets all constraints. Also, each individual member from the matched sample corresponds to a member from the treatment (or magnet) group. The group-based matching procedure introduced previously will often generate a matched sample with too little variability of the pretest scores, compared to the variability occurred in the treatment group. The individual matching procedure introduced in this section will alleviate this problem to some extent. The application of this matching procedure can be found on two program evaluation studies (e.g., Modarresi, Yang, Bulgakov-Cooke, & Li, 2004; Yang, Li, Modarresi & Tompkins, 2004). The other procedure to be introduced next will make not only the means of the pretest score but also their respective variances between two treatment and non-treatment groups very similar.

V. Multiple-Matched-Sample Procedure

A. Matching Procedure with the Involvement of Measurement Error

As indicated in the section of (group or individual) matching control, a unique matched sample will be generated once the criteria used for the matching procedure is determined. This is especially true if we assume the average pretest score of a treatment group (e.g., magnet program students) is a true score, not contaminated with any measurement error. For large sample sizes, this assumption should be appropriate. However, to increase the confidence level of seeking an appropriate matched sample as similar to the treatment group as we could obtain, such no-measurement-error is not necessary to be presumed by allowing the pretest-score mean to be contaminated with a "reasonable" measurement error. Equation 6 presented below will help us comprehend this concept.

$$\text{Minimize } \sum_{i=1}^{n}\left[\text{ABS}(P_i - (M + E)\right] x_i \qquad (6)$$

The components in Equation 6 are the same as those found in Equation 1, except the additional component of measurement error, E. The value of E can be randomly generated from the normal distribution, $N(0, SE^2)$, where SE represents the standard error of the mean of pretest scores for the treatment group. Specifically,

$$SE^2 = \frac{S^2}{N} \qquad (7)$$

Where

N is the sample size of treatment group,

S^2 is sample variance of pretest scores for the treatment group.

In reality, the matching procedure introduced in this section may still generate a matched sample whose pretest-score variance is still way off (too small) to the one found in the treatment group when the sample size of the treatment group, N, is too large. On the other hand, this matching procedure may generate a matched sample whose pretest-score mean is not very close

to the one found in the treatment group when the sample size is too small. This issue can be manually resolved by: first conducting several trials for various sizes of N; second, finding an appropriate N that will produce a matched sample whose mean as well as variance of the pretest score is very similar to the treatment group. Of course, another comprehensive approach can also be used for resolving this issue: first, decide how close the mean and variance of the pretest scores the two groups should be; second, repeatedly conduct the matching procedure until the criteria we set has been achieved. The iterative procedure that was often used in computer language (e.g., loops) can deal with this comprehensive approach very efficiently.

B. Multiple Matched Samples

By allowing the addition of measurement error into the mean score of the pretest for the treatment group during the process of matching procedures, a matched sample will be generated. Afterwards, every member of the non-treatment group is returned to the dataset after sampling. Another matched sample will be generated given a different value of measurement error. Again, every member of the non-treatment group should be returned to the dataset after sampling. After repeating the matched procedure again and again, multiple matched samples will be created. It is noted that many members from the population of the non-treatment group could appear multiple times in different sets of matched samples because the same constraints have been repeatedly imposed into the matching procedure.

The multiple-matched sample procedure creates a condition that the treatment group has multiple matched samples to compare with. As performed in the one-matched-sample approach, an effect size measure (for the discussions of the features of this measure, refer to Thompson [2002]) can be performed for each analysis in each comparison. If 1000 matched samples are used as in this evaluation, the mean as well as the distribution of the effect size measure, across 1000 replicated comparisons, can be used to assess the efficacy of any program. This enhances our confidence level to decide whether a program is effective or not.

A summative evaluation of the Reading Recovery Program employed this multiple-matched sample procedure to investigate whether or not students enrolled in the Reading Recovery Program gained any achievement advantage over students who were not enrolled in this program

(Yang, Li, Modarresi & Tompkins, 2004). Five hundred replicated matched samples were used in this evaluation. Figure 2, shown below, was the distribution of those effect size measures, across 500 replicated comparisons. The negative effect results indicated in this distribution seemed to suggest that the Reading Recovery Program did not raise reading performance of Reading Recovery students, compared with similar groups of students.

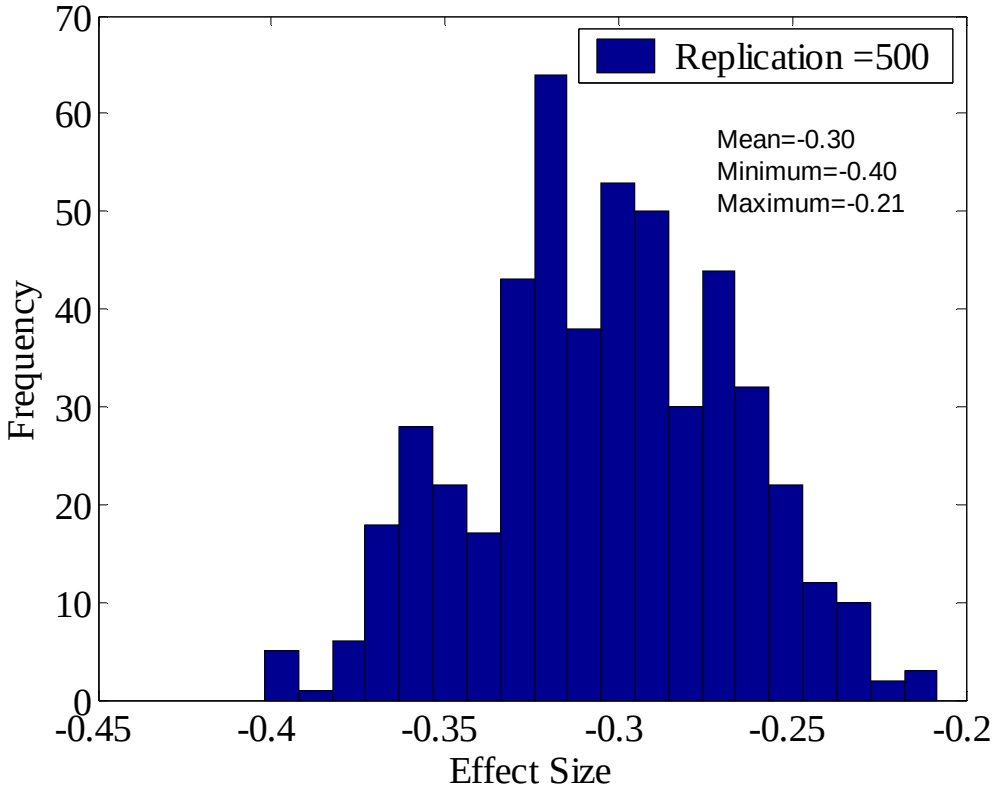

Figure 2. The Frequency of Effect Sizes for the Reading Recovery Group

VI. Discussions and Conclusions

A. The Solution of the *Zero-One Linear Programming*

In reality, the group matching procedure may be done using linear programming software packages (e.g., LINGO, see http://www.lindo.com) by "feeding" the **A** matrix and b vector into those computer software programs. However, if the problem is too complicated as usually

occurred in the matched procedure, it may not be very feasible to do so. For the individual matching procedure introduced in this paper, it almost cannot be done by directly using the software packages. Users need to write computer codes (e.g., C^{++} language, or MATLAB, The MathWorks, Inc., 2003) to call the callable libraries (e.g., LINDO API, LINDO Systems, Inc. 2003) to do so. For the solution presented in Table 3, the LINDO API was called into the MATLAB to seek the solution of the vector of $\underline{x}$ in Equation 4. All solutions met all the constraints without any difficulties, as well as in a timely manner (e.g. less than a second per matching).

B. The Features of the *Zero-One Linear Programming* Matching Method

Up to this point, this paper has introduced the use of the *Zero-One Linear Programming* in the context of quasi-experimental design to serve the specific purpose of creating a matched sample that is as similar to the experimental group as we can obtain. Comparatively, the propensity-score method is easier to implement in all types of quasi-experimental designs. However, using the method introduced here, researchers do not need to face tough issues, for example: Does the statistical model used for computing the propensity scores fit the data well?; Is the distributional assumption required for the statistical model violated? Furthermore, the propensity-score method is not suitable for the quasi-experimental designs required to have the pretest score information because this method does not handle such a covariate as appropriately as the one introduced here.

For the matched-sample introduced here, a <u>unique</u> matched sample will be generated once the criteria for the matching variables are determined and the measurement error on the pretest score is ignored. In addition, the identical distribution of different types of students (illustrated above) between the quasi-experimental and matched samples is a promising feature that cannot be found in pre-existing matching methods (e.g., the propensity-score method).

C. Statistical Modeling Followed by the Matching Method

<u>Evaluation for only One Program</u>

After the matching procedure, a small pretest score difference between the treatment group and its matched sample remained. The ANCOVA can be used to control for the effects of

the small pretest score difference. When the matching control is integrated with the ANCOVA analysis, ANCOVA resulted in *adjusted posttest means* for both groups under the constraint of two groups' pretest means being equal, as well as two groups' matching variables and the combinations of matching variables being equal. The latter constraint makes the ANCOVA-based adjusted means more defensible. If the data structure is hierarchical, the HLM model is preferred.

When the measurement error on the pretest score is taken into account, multiple matched samples can be generated and the performance of the treatment group can be compared with each of the multiple matched samples using the ANOVA analysis. The distribution (e.g., mean, minimum, and maximum values) of the effect size measure, across multiple replicated comparisons, can then be computed and used to assess the efficacy of any program. This enhances our confidence level to decide whether a program is effective or not.

Evaluation for Multiple Programs

When multiple programs or schools (e.g., 30 schools) are evaluated simultaneously, multiple matched samples will be generated --- each of matched samples will serve as a control group for a specific program. Under the logic of the randomized block design (Kirk, 1995), since students' outcome scores are more likely to be homogeneous within each program than across programs, the data from each program and its matched sample can be treated as a block. Within each block are students that received a program treatment or those that received no program treatment. The data from all blocks can then be aggregated to form a factor called "block" in the statistical context. Randomized Block Design assumes that unit (or student) assignment into both the treatment and non-treatment groups is random within each block. Since student assignment was not random in this study, however, the use of matched samples for the non-treatment groups represented an attempt to correct for this problem. This quasi-randomized procedure will make the adjusted-mean difference between two groups (the experimental group and control group) interpretable. The combined use of ANCOVA and Quasi-Randomized-Block Design was designed to reduce the error variance so that a more precise estimate of a treatment effect could be obtained. However, if the pretest-test means differ widely in different programs, using the ANCOVA to analyze each program's data separately, instead of using this combined

method to analyze all program's data simultaneously, is preferred because a separate ANCOVA analysis for each pair of treatment-and-matched data can avoid the use of the extreme extrapolation.

This combined method was used to investigate the effects of magnet school programs on the reading and mathematics performance of students in a school system (Yang, Li, Modarresi and Tompkins, 2005), in which the data collected from eight magnet programs were simultaneously analyzed and the effect sizes were then simultaneously computed for eight programs. Of course, HLM is another preferred method to model this type of data if the data structure meets the HLM requirements.

D. Concerns of Matched Samples

The relative success in creating a matched sample relies on which variables to base the matching as well as the selection of pretest scores to be minimized. In general, when more demographic variables are used in the process of matching, the result of the matched group's background is more similar to the experimental group; however, the gap of pretest score difference between the two groups might increase as the use of matching variables increases. This issue might be resolved by trying different combinations of matching variables and to see how large differences in the pretest score vary under different conditions. Based on those trial results, researchers then choose one suitable solution that fits their research interest the best.

In addition, the degree of successfully creating a matched sample also relies on whether the distributions of both the quasi-experimental group and its respective non-treatment group on the matching variables (especially on the pretest scores) substantially overlap or not. If both groups have more overlapping distributions on those matching variables, then the matched sample can be adequately obtained without the need of selecting members from extreme tails of the distributions. For example, the Non-Magnet population might have more overlapping distributions if such a population is composed of more members who are eligible for a specific magnet program, but they are not placed in this magnet program due to some circumstances (e.g., schedule conflict, no intention to attend, etc.). In contrast, the Non-Magnet population might have less overlapping distributions if such a population is only composed of members who are not eligible at all for this magnet program. When the later scenario occurs, examination

of the overlap of the two distributions will help alert researchers to the possibility of the regression effect among the matches (Shadish, Cook, & Campbell, 2002, p 121).

Finally, Shadish et. al. (2002) pointed out that matching can be done only on observed measures, so hidden bias may remain. Researchers should always be aware that in drawing conclusions from quasi-experimental designs, incorporated with the matched-sample method, causality may not be inferred due to the lack of random assignment of students to the treatment and matched groups. Only through random assignment of subjects can the two groups of subjects be equal on all possible observed and "hidden" variables. Of course, without a large sample, bias may remain even though random assignment is fully implemented.

References

Adcock, E. P. & Phillips., G. W. (2000, April). Accountability evaluation of magnet school programs: A value-added model approach. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models: Applications and Data Analysis Methods. Newbury Park, CA: Sage Publications, Inc.

CTBS/McGraw-Hill (1997). Teacher's guide to TerraNova. Monterey, CA. McGraw-Hill Companies, Inc.

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

Isaac, S. & Michael, W. (1995). *Handbook in research and evaluation*, (3nd Ed.). EdITS / Educational and Industrial Testing Service, C.A.

Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). Research Methods in Social Relations. San Francisco: Holt, Rinehart, and Winston, Inc.

Li, Y. H. & Schafer, W. D. (2005a). Increasing the homogeneity of CAT's Item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests, *Journal of Educational Measurement*.

Li, Y. H. & Schafer, W. D. (2005b). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. Applied Psychological Measurement, 29, 1-23.

LINDO Systems, Inc.,(2003). *LINDO API: User's Manual*. LINDO Systems, Inc, Chicago, IL.

Modarresi, S., Yang, Y. N. & Bulgakov-Cooke, D.& Li (2004, November). An investigation of the effects of an Algebra intervention program, *PLATO*, on the Algebra performance of students. Paper presented at the annual meeting of American Evaluation of Association, Atlanta, GA, November, 2004.

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 incorporate the propensity score. The American Statistician, 39, 33-38.

Rosenbaum, P. R. (1995). Quantiles in nonrandom samples and observational studies. Journal of the American Statistical Association, 90, 1424-1431.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. Journal of Educational and Behavioral Statistics, 29(1), 103-116.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton Mifflin Company, M.A.: Boston.

The MathWorks, Inc. (2003). MATLAB (Version 6.5): The language of technical computing [Computer program]. Natick MA: The MathWorks, Inc.

Theunissen, T. J. J. M. (1985). Binary programming and test design, *Psychometrika*, 50, 411-420.

Theunissen, T. J. J. M. (1986). Optimization algorithms in test design, *Applied Psychological Measurement*, 10, 381-389.

Thompson, B. (2002). "Statistical," "Practical," and "Clinical": How many kinds of significance do counselors need to consider?

van der Linden, W. J., & Boekkooi-Timminga, E. (1989).A maximin model for test design with practical constraints, *Psychometrika*, 54, 237-247.

Yang, Y.N., Li, Y. H., Modarresi, S., & Tompkins, L., J. (2003). An investigation of the effects of magnet school programs on the Reading and Mathematics performance of students. Prince George's County Public Schools, Maryland.

Yang, Y.N., Li, Y. H., Modarresi, S., & Tompkins, L., J. (2004). An investigation of the maintenance effect for first graders enrolled in the Reading Recovery program on their succeeding grade two Reading performance. Prince George's County Public Schools, Maryland.