An Research Methodology for Future Summative Evaluation Studies: Incorporating the Component of Multiple Sets of Matched Samples into the Statistical Control Modeling

By

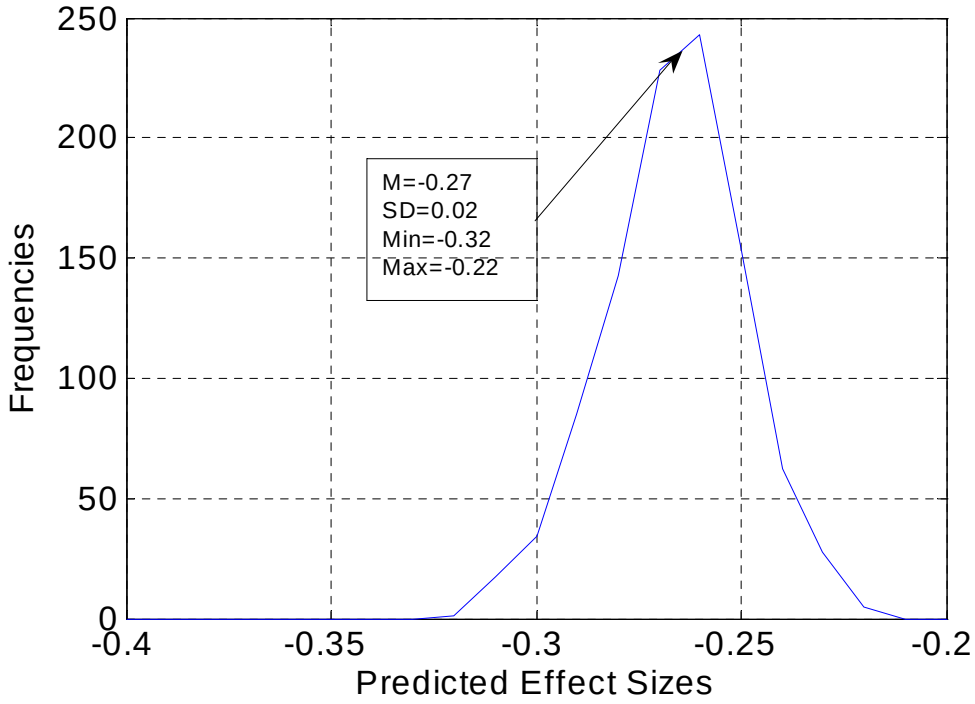
Yuan H. Li, Ph.D.

Prince George's County Public Schools, Maryland

Shahpar Modarresi, Ph.D.

Montgomery County Public Schools, Maryland

Yu, N. Yang, Ph. D.
Prince George's County Public Schools, Maryland



The Distribution of the 1000-Replicated Effect Sizes for Estimating the Effects Sizes

Paper was presented at the annual meeting of the American Educational Research Association, San Francisco, CA, April, 2006.

A Research Methodology for Future Summative Evaluation Studies: Incorporating the Component of Multiple Sets of Matched Samples into the Statistical Control Modeling

ABSTRACT

Summative evaluations have often been undertaken to determine the impact of educational programs on student academic achievement employing a quasi-experimental design. The summative finding is expected to be less misleading if a statistical model is performed on a dataset including a sound matched sample as a control group. This is because an extreme or untrustworthy extrapolation is not necessary to be applied and a regression artifact will be migrated. Empirical results showed that with the use of a single set of matched samples, the estimated program's effect might be unstable, however. A research methodology incorporating the components of multiple sets of matched samples in addition to the statistical control modeling (e.g., ANCOVA) is expected to mitigate this problem because this proposed method is expected to consistently reduce the selection bias in the quasi-experimental designs since almost all possible sets (e.g., 1000) of matched samples are drawn from the non-treated population.

The proposed multiple-sets-of-matched-samples method creates a condition in which a single treatment group has multiple sets of matched samples as control groups. If 1000 matched samples are drawn from the non-treated population and the effect size (ES) measure is then estimated for each of 1000 comparisons, the mean as well as the distribution of those 1000 ES measures can be a more reliable measure to assess the efficacy of educational programs. This enhances our confidence level to determine whether or not a program is effective.

A summative evaluation study of a technology-based reading intervention program that has utilized this research methodology was introduced in this paper to enunciate the appealing features of this research methodology.

Index: Matched Sample, Matched and Statistical Control, Zero-One Linear Programming, ANCOVA, Propensity Score Matching, Summative Evaluation.

I. Introduction

A. Background of Program Evaluation Designs

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 (or covariate) variables because it makes both groups very similar at the beginning of an experiment, especially in cases where the sample size is large. Unfortunately, random assignment is not often feasible in educational settings due to ethical, practical, and logistic issues. School administrators commonly believe that it is unethical to deny a potentially beneficial program to students who need the program merely for evaluation purposes. There are many practical and logistical issues that may also occur during the course of the study. For example, a student assigned to a treatment group may transfer to another classrooms, school or school system. There is also the problem of mortality. Accordingly, quasi-experimental designs (Isaac & Michael, 1995; Shadish, Cook & Campbell, 2002) become one of the alternatives that may be used to determine a program effect in educational settings.

Among the quasi-experimental designs, the *non-randomized comparison group pretest-posttest design* (illustrated in Figure 1, refer to Shadish, Cook & Campbell [2002], p. 136) is one of the sound evaluation designs in assessing the efficacy of any programs. In this design, random assignment is not conducted and subjects in both the quasi-experimental and the control groups take both the pretest and the posttest. The pretest is often used to measure subjects' initial ability and the outcome measure is often assessed by 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). The number of the quasi-experimental group could be single (e.g., only one program) or multiple (e.g., several programs to be evaluated simultaneously).

	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. The Need for the Utilization of Matched Samples

In gauging the efficacy of any programs that have incorporated the evaluation design as shown in Figure 1, the analysis of covariance (ANCOVA, Kirk, 1995) is often used to account for individual students' characteristic difference (e.g., sex, race, pretest scores, etc.) and the hierarchical linear modeling (HLM, Raudenbush & Bryk 2002) is often used to account for both individual and organization (e.g. school) context differences (e.g., percent of minority students, percent of poverty students, etc.) between the quasi-experimental and control groups.

However, as pointed out by Rubin, Stuart and Zanutto (2004), comparing results obtained from treated and all non-treated student population with very different distributions of background covariates will heavily rely on statistical modeling assumptions (e.g., linearity assumption) as well as extreme extrapolation. Reliable causal inferences may thus not be drawn. For example, Rubin et al. (2004) pointed out that the values of "percent minority" and "percent in poverty" may differ widely in some schools, and this situation will cause the estimated school effects that have been adjusted for such covariates to be extremely sensitive to the statistical modeling assumptions. If the assumptions are seriously violated, those estimated school or program effects, as a result of using extreme or untrustworthy extrapolation and/or a regression artifact, would be seriously misleading.

Instead of using all non-treated student population as a control group, drawing a sound matched sample from the non-treated student population as a control group is expected to be a more appropriate step before any statistical modeling is performed. A set of matched samples helps ensure that the statistical modeling is done using treated and control groups with similar covariate distributions, thus reducing reliance on the linearity assumption and

extreme extrapolation. Nevertheless, how to create an appropriate matched sample for each program evaluation is a challenging issue that is reviewed below.

C. Matching Via the Use of Propensity Scores

If only a small number of categorical variables (e.g., gender and poverty status) and one or two continuous variables (e.g., pretest scores) are considered for matching, an exact matching method can be easily done. The procedure creates exact match on the categorical covariates (e.g., gender and poverty status) and then chooses the closest matches on the continuous variable (e.g., pretest score or any other covariate) within those exact matches. This matching method, similar to the greedy matching method, will seek the closest control match for each treated sample one at a time, without trying to minimize a global distance measure. The result of this matching cannot ensure that the criterion set by the optimal matching method will be met, where the optimal matching will find the matched samples with the smallest average absolute distance across all the matched pairs.

In instances where researchers attempt to balance groups simultaneously on many covariates (e.g., on gender, ethnic groups, poverty status, pretest score, the total days that students attended at their schools, etc.), the above exact matching method may encounter practical problems. Diverse methods can be used to serve this purpose (for literature review, see Rosenbaum, 2002). Among them, Rosenbaum and Rubin (1983, 1984, 1985) proposed an approach that involves the use of propensity score. The propensity score is a single composite score that represents a person's score on all observed covariates and it is often estimated by the logistic regression modeling. Once the propensity score is estimated, it can be matched, blocked in quintiles (or stratification), used as an ANCOVA covariance adjustment, or weighted based on researchers' interest. The goal of propensity score analysis is to balance two nonequivalent groups on observed covariates in order to get more accurate estimates of the effects of a treatment.

Numerous published papers on propensity score matching exist (e.g., Rosenbaum & Rubin 1983 and 1985; Rubin and Thomas, 1992a and 1996). Propensity score has become popular for matching in quasi-experimental designs. However, this popularity is only warranted if the propensity score is accurately estimated. The fundamental assumption underlying matching estimators is that all variables related to both outcomes and the

treatment assignment variables are included in the vector of observed covariates (Rosenbaum & Rubin, 1983). This assumption emphasizes that researchers fully know what the correct model is. In reality, researchers don't fully know the correct model and need to empirically seek that true model. The issues of model selections (e.g., inclusion or deletion of covariates, the interactions among covariates, the higher-order terms of covariates, etc.) will be the first task to be adequately addressed. Furthermore, the aptness of the selected logistic regression model needs to be examined before it is accepted for use, as is the case for all regression models. In particular, researchers need to examine key properties of the logistic response function. This means that researchers need to determine whether the estimated response function for the data is monotonic and sigmoidal in shape.

In short, before using the propensity score, the issues indicated above should be closely examined to ensure the meaning of the propensity scores as the propensity-score theory has stated (Rosenbaum & Rubin 1983). Research on how well the theoretical result is satisfied when using estimated rather then true propensity score can be found on several studies (e.g. Rubin & Thomas, 1992b, 1996). Although the estimated true score generally worked well as shown in some studies (e.g. Rubin & Thomas, 1992b, 1996), those findings may not be generalized to all research circumstances if the issues indicated above (e.g., model selection, model-data fit, etc.) are on purpose ignored.

D. The Need of Matching on Covariates

As indicated previously, balancing groups simultaneously on many covariates is a tedious task. A composite index of the propensity score for all types of covariates was originally designed to simplify this process. It seemed that there was no more need for direct matching on those observed covariates themselves. Nevertheless, a simple propensity score model may not fully account for the importance of those more important covariates (e.g., the pretest-test score) in terms of predicting the outcome of interest. In the logistic model, 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). As a matter of fact, some covariates (e.g., the pretest score) are usually highly correlated with the outcome measure rather than the treatment assignment. This fact

may cause the pretest-score covariate to be less important than it should be when only the propensity-score is used for matching.

To address the issue indicated above, several modified propensity-score matching methods did not totally rely on the propensity score for matching but also rely on the observed covariate variables. For example, Rubin and Thomas (2000) proposed a method by matching individual units on some measure of distance between their covariates within strata of propensity scores. Using this matching method, the importance of the variables (e.g., the pretest score) that predicts most of the variance of the outcome measure can be addressed if they are chosen as key matching variables. Other proposed matching methods also include the covariates for matching, for example, one matching method that combines the Mahalanobis metric method with the propensity score matching (e.g., refer to Rosenbaum & Rubin, 1985).

In practice, using the matching procedures that totally rely on the propensity scores may end up with a male African-American student matched with a female White student because both students might have very similar propensity scores. This matching method can be analogous to a "Fruit Smoothie by Fruit Smoothie" matching method because researchers put all kinds of covariates (or fruits) into a statistical model (or a smoothie blender) to produce a composite score (or fruit smoothie). If the propensity scores are well estimated, the bias created by nonrandom assignment, in theory, is expected to be reduced by the use of the propensity scores (Rosenbaum & Rubin, 1983). However, a practical issue may occur, that is, decision makers, school personnel (e.g., principals) or general public may have a hard time to accept that the samples (treatment and control) are similar.

Due to other rationales (e.g., effectively reducing selection bias) or/and the practical issue indicated above; researchers may not only rely on the propensity score for matching but also include several key covariates (e.g. pretest score) in the matching procedures. This issue suggests that the central benefit of the propensity score —simplifying the matching process by reducing multiple dimensions to one has become less important than it was originally developed. Several modified-version of the propensity score matching methods prefer not only the propensity score but also the covariate variables for matching.

In the past, without the sophisticated technical support, matching on several covariates was a tedious task or even impossible if both the greedy and optimal matching criteria were required to be met. With the combination of advances in the computing power of personal computers and the applications of the zero-one linear programming (e.g., Li & Schafer, 2005a; Li & Schafer, 2005b; Li, Yang & Modarresi, 2005), the tedious task in matching several covariates has disappeared and a set or even multiple sets (e.g., 1000) of matched samples can be created in a timely manner. Specifically, without any preference for the statistical model selection as often required by the family of the propensity score matching methods, the matching procedure developed by Li et. al. (2005) is very efficient in matching as many categorical covariates (e.g., gender, ethnic) and continuous covariates (e.g., the amount of family income, total days in a school) as the researcher desires. The criterion of the exact matching can be achieved either in individual or subgroups for the categorical covariates and the criterion of optimal matching can be achieved for the continuous variables. Furthermore, the proposed matching method can appropriately handle the key covariate (e.g., pretest score) simultaneously using the greedy and optimal matching methods.

To sum it up, with the aid of the *Zero-One Linear Programming* and the Li et al.' matching algorithms (2005), the balance of groups simultaneously on many covariates becomes possible. The greedy and optimal matching methods can be applied on the key covariate (e.g., pretest score). The exact matching method can be applied on categorical covariates and the optimal matching method can be applied on continuous covariates. This can be done without the need of using a statistical model to generate a composite score. This matching can be analogous to an "Apple by Apple" matching method because all kinds of covariates (or fruits) remain their original forms during the matching process. This "Apple by Apple" matching method attempts to ensure both the "Chemicals" and " Real Persons" between the quasi-experimental and matched groups are very similar. The property of matching both "Chemicals' and "Real Persons" simultaneously is appealing in particular when a researcher explains to non-technical persons how similar both the treatment and matched groups are.

E. The Needs for Creating Multiple Sets of Matched Samples

As stated above, it is expected to obtain more reliable and accurate evaluation results when matching strategies are used to create a control group and consequently a statistical model (e.g., ANCOVA) is applied to the data. Nevertheless, based on the findings from an empirical study (Yang, Li &Modarresi, 2005), the results produced by the single-matched-sample approach might not be adequately reliable to make a right decision especially in the cases with a small sample size (e.g., 30, 50). For example, the effect size measure for one of the program evaluation was about 0.27 in a study (Yang, Li, Modarresi & Tompkins, 2003) conducted by a single matched-sample method. At that time, this program was not phased out by the decision makers partly because this effect size value was larger than the criterion of 0.2 set by that evaluation study.

Nevertheless, when the 200 sets matched samples were drawn from the same sets of data and they were then re-analyzed, the effect size values ranged from –0.47 to 0.29 with a mean of –0.16 (Yang, Li & Modarresi, 2005). The latter finding contradicts the original finding and suggests that parameter estimate calculated by using a single matched sample for determining a program's effectiveness may be unstable. This seems to suggest that using a single set of matched samples cannot ensure the reduction of the selection bias to the level that most researchers are confident of it. Accordingly, the mechanism (or algorithms) for generating multiple sets of matched samples and using them in quasi-experiment designs deemed to be necessary in order to consistently control the selection bias. This will consequently obtain a stable estimate of program effectiveness.

The matching method proposed by Li et al. (2005) creates multiple sets of matched samples that are then treated as replicated-similar multiple control groups. This feature is another attribute that has not been addressed by the current matching methods in the literature . As delineated in the Li et al.'s study (2005), the amount of measurement error of the pretest score (one of the most important covariates) was incorporated into the equation used for creating multiple sets of matched samples. The proposed multiple-matched-sample method is expected to "consistently" mitigate the problem of selection bias in the quasi-experimental designs ------ since almost all possible sets (e.g., 1000) of matched samples are drawn from the non-treatment population. Of course, this method as the same as other

matching methods will work only the important covariates are sought and used in the matching process.

Each set of control groups is similar to the treatment group and can be used for the purpose of comparison. In creating a single set of matched sample, the principle of sampling without replacement is applied. Any non-treatment member, thus, can not appear multiple times in a set of matched sample.

After sampling a set of matched samples, every member of the non-treatment group is returned to the dataset. Another new set of matched samples will be generated given a different value of measurement error of the pretest score to each member in the quasi-experimental group. Again, every member of the non-treatment group should be returned to the dataset after sampling. After repeating the matching procedure again and again, multiple sets of matched samples will be created. As a result of the procedures used to generate multiple sets of matched samples, many members from the population of the non-treatment group could appear multiple times in different sets of matched samples, but not within any sets of matched sample. The reason is that the same constraints have been repeatedly imposed while creating any sets of matched samples.

Confidence interval of program effectiveness can be established by the percentile method based on the distribution of multiple replicated effect size estimates. This cannot be found in the relevant literature.

II. Purpose

Once the multiple sets of matched samples are drawn from the non-treated population, this creates a condition that the treatment group has multiple matched samples for the purpose of comparison. An effect size measure (ES, Thompson, 2002) can be calculated for each of the comparisons. If 1000 matched samples are used, as is the case in this evaluation, the mean as well as the distribution of the effect size measures, across 1000 replicated comparisons, can be used to assess the efficacy of any program. The use of this method enhances our confidence in determining whether or not a program is effective.

The incorporation of the components of multiple sets of matched samples into the statistical modeling creates a research methodology for future summative evaluation studies.

This research methodology is a contribution to the literature and is expected to be more reliable to reduce the selection bias when compared to a single-matched approach.

This paper attempted to enunciate the appealing features of this proposed research methodology by using a summative evaluation study as an example. The summative study was undertaken to determine whether or not attending a technology-based reading intervention program leads to a higher student achievement. This summative question was addressed using the proposed research methodology. The reading performance of students who participated in the stated program and received treatment was compared to 1000 sets of matched samples of non-treatment peers. The mean and the distribution of the 1000 effect size measures were used to assess the program's effectiveness. The details of this summative evaluation are briefly illustrated below.

III. A Summative Evaluation of a Reading Intervention Program

A. Program Descriptions

A technology-based reading intervention program is designed to meet the needs of students whose reading achievement is below their grade-level proficiency. The developers of this program believe that research on the causes for poor reading performance for elementary and secondary students rests on two areas: 1) poor decoding and slow oral reading fluency; and 2) difficulty in creating mental models from the text.

The solution to these problems is to provide an instructional setting that has a small class size, extended teaching time devoted to reading, writing instruction, and the use of differentiated instructional strategies by the teacher. The use of technology during the small group rotation part in the reading intervention lesson allows students to become more accurate, automatic, and fluent readers. During the small group teacher directs mini-lesson instruction, students receive differentiated instruction based upon their strengths and weaknesses in comprehension. This program incorporates technology in learning that has strong proponents in public education. This program was administered to ninth graders at five high schools in a school district. A summative evaluation was conducted and the key question for this evaluation is described below.

B. Research Questions

The major summative question for the evaluation of the technology-based reading intervention program is stated below:

Do students in the technology-based reading intervention perform better in reading than similar non-reading intervention program after controlling for students' demographics (viz., poverty, race, and gender) and their initial abilities?

C. Research methodology

1. Evaluation Design

A Quasi-experimental design was used in this evaluation to address the evaluation question. Multiple sets of matched samples were drawn from the non-treated student population, at the five reading intervention high schools, for the purpose of comparisons. Figure 2 illustrates the evaluation design, where it shows that non-reading intervention students took regular English 9 course; whereas, reading intervention students not only attended the regular English 9 course but also received the reading intervention treatment. In other words, the reading intervention program was an additional treatment for those selected high school students who were enrolled in the program.

The 2003 Maryland School Assessment (MSA) reading and 2004 High School Assessment (HSA) reading tests were used as pretest and posttest measures, respectively.

	Pretest		Treatment		Posttest
Quasi-Experimental Group (Reading Intervention Program Students)	O_1	=>	X	=>	O_2
Control Group (1000 Sets of Matched Samples)	O_1	=>	C	=>	O_2

O_1 – 2003 MSA Reading Scores for eighth graders
C – Taking English Course
X – Taking English Course + the Reading Intervention Program treatment
O_2 – June 2004 HSA English Score for ninth graders

Figure 2: The Summative Evaluation Design of Reading Intervention Program

2. Matching Control

Four categorical covariates were considered for matching for this evaluation. They were, gender (2 levels), race (5 levels), poverty (3 levels), and school (5 schools) that students attended. Under this circumstance, there are 150 (2x5x3x5=150) types of students. In addition, the pretest score and the total days that students attended at the school were considered for matching. There were 155 and 1505 students in the reading intervention program and non-reading intervention programs, respectively. One thousand sets of matched samples were drawn from the non-treated student population. As explained in the Appendix A, many members from the population of the non-treatment students could appear multiple times in different sets of matched samples because the same constraints, as described below, have been repeatedly imposed into the matching procedures.

For each set of the matched samples, 155 reading intervention students were drawn from the 1505 non-reading intervention program students with the following constraints:

1. The average pretest score of the reading intervention program group was close to the average pretest score of the non-reading intervention matched group.
2. Each individual pretest score of the reading intervention program sample was close to pretest scores of the reading intervention matched sample.
3. Each individual reading intervention sample had the same demographic characteristics (e.g., race, gender, and poverty status — Free/Reduced/Paid Lunch) as the reading intervention program matched sample.
4. Each individual in the reading intervention sample attended the same high school as the one in the matched sample.
5. The total days (TD) that the reading intervention program group attended at school was similar to the total days attended by the non-reading intervention matched group. Specifically, the following constraint was imposed: The TD for the reading intervention group can't be smaller than its TD minus one TD's standard deviation and its TD can't be larger than its TD plus one TD's standard deviation.

When those constraints were imposed in the zero-one linear programming, sets of matched samples were created. Thirty out of the 155 pairs are shown in Table 1. For each pair matching, several matching variables (e.g. the pretest score, total days in schools, sex,

race, poverty status, types of schools) are used and shown in Table 1. The definitions of labels in Table 1 are provided in the footnote.

The descriptive statistics of the pretest scale scores and the total days students attended at schools for the reading intervention group and matched sample groups are presented in Table 2.

As shown in Table 1, most members in the quasi-experimental group found respective matched samples with almost the same pretest scores. The results in Table 2 reveal that the summary statistics, Means, Standard Deviations (SD), Minimum (Min), and Maximum (Max), for the pretest scores were almost the same for the two groups of students (treated and non-treated). The combination of findings presented in Tables 1 and 2 reveals that the criteria set by the greedy and optimal matching methods have been achieved for the pretest score variable.

For the continuous covariate variable-- the total days students attended at schools-- we also found that each pair has a similar number of the total days attended at schools even though the greedy (or one-by-one) matching method was not implemented on this variable. The results in Table 2 show that the summary statistics, Mean, Standard Deviation (SD), Minimum (Min), and Maximum (Max), for this covariate are almost the same between the two groups.

It is important to note that "the same demographic characteristics" on the matching variables (e.g., poverty, gender.) not only means that the number of students on each selected variable is identical between the two groups, but also the distribution of students in the combination (e.g., types of students) of matching variables (e.g., poverty and gender) is identical. The above stated 5 constraints make the matched control samples generated by this matching method as similar to the reading intervention samples as they could be. Of course, if researchers can collect information on more key covariates, they may add those to the existing matching variables using the matching method introduced in the Appendix A.

Table 1:
Thirty Matched Pairs on Several Matching Variables (e.g. the pretest scores, days of students attending at the school, sex, race, poverty status, attending school and the types of student) Using the Matching Method Proposed by this Research

Pair	Type		Race		Gender		Poverty		School		Pretest		Days in Schools	
	Pro	Mat	Pro	Mat	Pro	Mat	Pro	Mat	Pro	Mat	Pro	Mat	Pro	Mat
1	3	3	3	3	1	1	1	1	1	1	354	356	171	153
2	3	3	3	3	1	1	1	1	1	1	352	354	175	174
3	3	3	3	3	1	1	1	1	1	1	379	383	147	167
4	3	3	3	3	1	1	1	1	1	1	365	365	148	173
5	3	3	3	3	1	1	1	1	1	1	405	406	146	163
6	3	3	3	3	1	1	1	1	1	1	386	386	162	178
7	3	3	3	3	1	1	1	1	1	1	390	394	152	176
8	3	3	3	3	1	1	1	1	1	1	390	390	147	177
9	3	3	3	3	1	1	1	1	1	1	363	363	157	172
10	3	3	3	3	1	1	1	1	1	1	351	351	137	149
11	3	3	3	3	1	1	1	1	1	1	379	376	157	141
12	5	5	5	5	1	1	1	1	1	1	401	402	165	159
13	5	5	5	5	1	1	1	1	1	1	383	390	154	162
14	6	6	3	3	2	2	1	1	1	1	362	360	171	152
15	6	6	3	3	2	2	1	1	1	1	360	360	166	170
16	6	6	3	3	2	2	1	1	1	1	356	356	155	169
17	6	6	3	3	2	2	1	1	1	1	342	342	156	134
18	6	6	3	3	2	2	1	1	1	1	367	367	166	143
19	6	6	3	3	2	2	1	1	1	1	362	363	154	162
20	6	6	3	3	1	1	1	1	2	2	383	383	178	157
21	6	6	3	3	1	1	1	1	2	2	383	383	177	156
22	6	6	3	3	1	1	1	1	2	2	381	381	170	163
23	6	6	3	3	1	1	1	1	2	2	362	362	160	173
24	9	9	3	3	1	1	3	3	1	1	356	356	174	176
25	9	9	3	3	1	1	3	3	1	1	377	377	175	173
26	9	9	3	3	1	1	3	3	1	1	370	372	158	154
27	9	9	3	3	1	1	3	3	1	1	406	405	161	165
28	9	9	3	3	1	1	3	3	1	1	374	374	153	166
29	9	9	3	3	1	1	3	3	1	1	384	384	169	176
30	9	9	3	3	1	1	3	3	1	1	381	381	163	125

Pro: Treatment Group; Mat: Matched Group; Race,1 for American Indian, 2 for Asian American, 3 for African American, 4 for White, 5 for Hispanic; Gender, 1 for Male, 2 for Female; Poverty: 1 for Free Lunch, 2 for Reduced Lunch, 3 for Self-Paid Lunch.

Table 2.
Descriptive Statistics of the Pretest Scale Scores and the Total Days Students Attended at Schools for the Reading Intervention and Matched Sample Groups

Group	N	Type of Students	Pretest Scale Score				Days of Student in the School			
			Mean	SD	Min	Max	Mean	SD	Min	Max
Treatment	155	The Same	371	19	311	418	167	10	129	178
Matched	155		372	18	317	418	166	11	123	178

The researchers also used the greedy matching method with the propensity score for selecting a set of matched sample (thirty out of 155 pairs are shown in Appendix B, Table B-1). The propensity score was estimated based on a logistic regression model with the status of students' treatment (Treatment/Non-treatment) as the dependent variable and the following variables as continuous and categorical covariates: Pretest scores, total days in schools, race, gender, poverty and school that student attended. No interaction variables among covariates were entered in the model.

As shown in Table B-1, each pair has almost the same propensity scores. Nevertheless, they are different type of students. For example, in the first pair, an Africa-American/Male/Free Lunch/School 3 student is matching with an Africa-American/Female/Reduced Lunch/School 3 student. The summary descriptive statistics of the pretest scale scores and the total days students attended at schools for the reading intervention and matched sample groups are presented in Table B-2.

3. Statistical Control

After the matching procedure, a small pretest score and other covariance differences between the program sample and its matched sample may remain. The analysis of covariance (ANCOVA) is used to control for the effects of these small differences between the two groups.

The application of ANCOVA procedures on the data containing a matched group results in *adjusted posttest means* that are more defensible. This is because an extreme or untrustworthy extrapolation is not necessary to be applied and a regression artifact will be migrated while estimating those adjusted posttest means.

4. Evaluation Criterion:

When a statistically significant result is obtained (e.g., $p < .05$), the researcher generally has the confidence to conclude that the treatment effect is not due to sampling error. A mean difference of a very small size can be judged to be statistically significant when the sampling error is small (due to a large sample size). Conversely, a relatively large mean difference can be judged to be not statistically significant when the sampling error is large (due to a small sample size). To avoid those problems above, this evaluation thus chose the effect size measure as a criterion to determine program effectiveness because such index resulting from the current evaluation design will allow for the isolation of the "program effect" (or the "value added") unencumbered by several key demographic factors and student's initial abilities. The effect size measures were calculated by the following procedures:

a. The Value-added Score:

The main task of program evaluation is to estimate the amount of the value-added scores of students that is contributed by the program. The definition of the value-added score is the change attributed by a specific experience (e.g., a reading intervention program treatment).

The *value-added score* (or non-standardized effect size) is mathematically defined as: Adjusted $\text{Mean}_{\text{Program Treatment}}$ minus Adjusted $\text{Mean}_{\text{MatchedSample}}$, where the adjusted means were estimated by the ANCOVA model for the reading intervention group and its matched-sample group. The value-added score was used to measure the magnitude of the reading intervention program effect.

b. Effect Size:

Since the value-added score depends on the scale score of the posttest and cannot be used to compare program effects among multiple programs, a standardized effect size (ES) with a promising feature of scale invariant or metric-free was used to compare the treatment effects among multiple school sites, subgroups, or multiple-replicated comparisons. Effect size, as illustrated in Equation 1 below, can be defined as the *valued-added score* divided by the standard deviation (SD) of the pooled posttest scores.

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

Cohen (1988) suggests 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. With 1000 ES values, a mean of effect size of 0.2 is adopted by this study to show efficacy of this reading intervention program. The interpretation of ES is provided in the Appendix C.

D. Summary of the Summative Evaluation Findings

After the students' characteristics and initial abilities (e.g., SY 2003 MSA pretest scale scores) were accounted for both the program's participants and the 1000 sets of similar non-reading intervention peer groups, the effect sizes were calculated and summarized in Table 3. The distribution of the 1000 effect size measures is shown in Figure 3.

Table 3.

The Summary Descriptive Statistics of the 1000-Replicated Effect Sizes Computed for the Reading Intervention Program on the SY 2004 HSA English Test Scores

Factors	Grade Level	Sample Size	Mean of ES	SD	Min	Max	PR# for the Mean ES
Whole Group	Nine	155	-0.27	0.20	-0.32	-0.22	39

#: The PR stands for the percentile rank (PR) standing of the treated sample mean when it is compared with the distribution of the matched-sample test scores.

The analysis of the 1000 sets of matched samples reveals that effect size (ES) values range from -0.32 to -0.22 with a mean of -0.27. The mean of the effect sizes is relatively smaller than the criterion of 0.2 . This would imply that a typical reading intervention student performed no better than a typical student who did not participate in this program.

The shape of the distribution of the 1000 effect sizes approximates a normal distribution (see Figure 3). This suggests that if a single-matched-sample method was

utilized to evaluate this reading intervention program, as found in most literature, program evaluators would obtain an ES value, ranging from –0.32 to -0.22. By using the multiple-sets-of-matched-samples method, the researchers have more confidence to conclude that the reading intervention program had minimal, if any, program effect on ninth grade students' English performance.

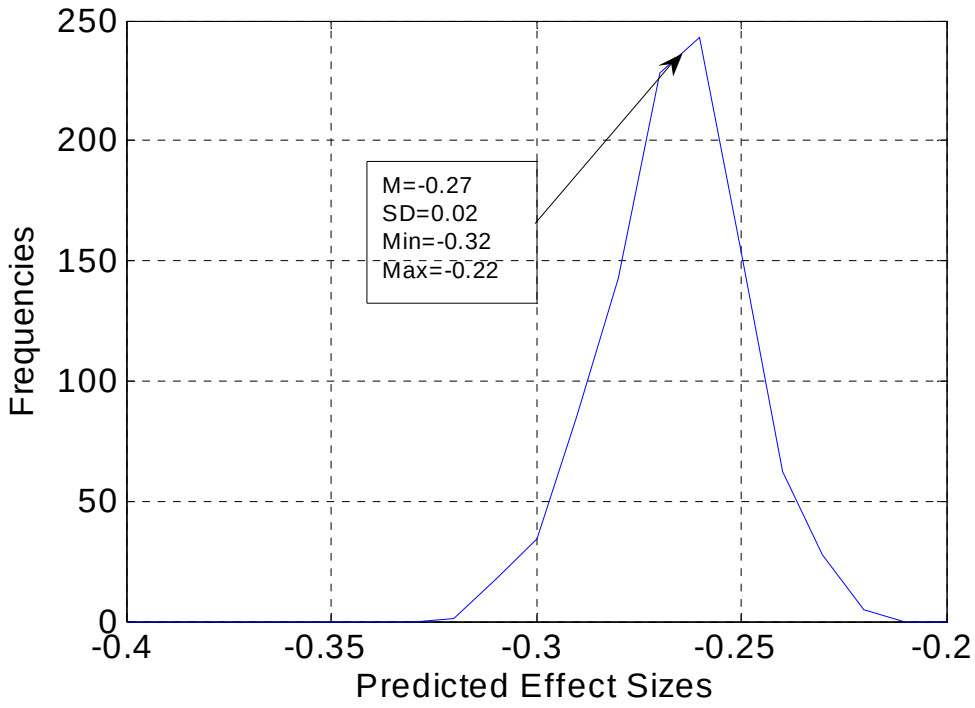


Figure 3. The Distribution of the 1000-Replicated Effect Sizes for Estimating the Effects of Reading Intervention Program

IV. Discussions

A. Features of the Proposed Research Methodology

The research methodology introduced in this paper heavily relies on the matching method illustrated in Appendix A. Readers might think that the matching introduced in the Appendix can be done by simply cross-classifying the control group on the categorical covariates, sorting the control group based on the order of distance of pretest score from a member's "pretest score" and then, within each cross-classification of the categorical covariates, choosing the matched sample who have the smallest distance to this member's "pretest score". Unfortunately, as stated previously, the stated matching method is a one-by-

one matching technique that can't minimize the pretest-score mean differences between the quasi-experimental group and the matched group.

Also, in using the above described cross-classification matching method, researchers may choose to sort the control group based on the order of distance of pretest score from the pretest mean score of the quasi-experimental group and then, within each cross-classification of the categorical covariates, choose j matched samples who have the smallest distance to the pretest mean score of the quasi-experimental group, where j is the number of quasi-experimental members in the cross-classification group. Unfortunately, this modification of the above cross-classification matching method is like an optimal matching method and does not ensure that individual units on the pretest scores are well matched.

Both cross-classification matching methods introduced above will encounter practical problems when covariates have many levels or when the number of covariates is large. In other words, using this cross-classifying method to balance groups simultaneously on many covariates is an arduous task. Furthermore, the criteria set by both greedy and optimal matching methods might not be met simultaneously. Fortunately, with the aid of zero-one linear programming and the algorithms introduced in Appendix A, the problem of tedious task on matching is solved. Furthermore, the following features will be detained:

(1) Greedy and Optimal Matching for the pretest scores (or other key covariates). This mean each member in the quasi-experimental group has a respective matched sample with almost the same pretest score. In addition, the mean of pretest scores between the quasi-experimental group and the set of matched samples is nearly the same.
(2) Identical (one-by-one) matching for the demographic/background variables between the treated and non-treated groups.
(3) An optimal matching method is applied for all continuous covariates.
(4) Taking into account the measurement error of the pretest scores while creating multiple sets of matched samples.

The tedious task on matching many covariates simultaneously can also be solved on matching the composite score of the propensity scores. However, despite the use of the propensity score for matching, researchers may still want to select several other key covariates for direct matching (e.g., Rubin & Thomas, 2000). Furthermore, the issues of

model selection and model-data fit cannot be ignored before the propensity score is used. In other words, in most cases the tedious work of matching is not fully solved regardless of using the propensity score. Therefore, the proposed matching method in creating multiple sets of matched sample can be one of new promising matching method.

Some researchers may ask the following question: Is the results produced by the single-matched-sample method reliable enough for making policy-decisions of phasing out or keeping a program, compared to the multiple-set-of-matched-sample approach? According to an empirical study (Yang, Li & Modarresi, 2005), the results produced by the single-matched-sample approach may not be reliable enough to make a correct decision especially for a case with a small sample size (e.g., 30, 50). As we are aware, sample sizes in many programs are small. The research methodology: Incorporating the Components of Multiple Sets of Matched Samples into the Statistical Control Modeling, increases our confidence in conclusions of summative evaluations of small programs.

In the evaluation of the reading intervention program, the measurement error of the key variable (pretest score) was also taken into account during the process of matching the variables. Figure 3 shows the distribution of the multiple effect size measures, where it shows the variability of effect size measures. Such ES variability is primarily introduced by the fluctuation of multiple samplings as well as the variation of the key covariate of the pretest-score measure. Such plot provides the richest information for decision makers and cannot be found in other literature that is associated with summative evaluations.

To date, simulation studies conducted to compare the performance among several matching methods are rare. Researchers should continue to work on this issue. However, before such information is available, the mechanism (or algorithm) to generate multiple sets of matched samples in summative evaluations using the quasi-experiment designs is essential, no matter what matching method the researchers prefer.

B. Issues Related to the Applications of this Research Methodology

Several factors associated with this research methodology are discussed below.

1. The Feasibility of this Research Methodology

In reality, the single matching method as well as the multiple-sets-of-matched-sample method cannot be done by directly entering the **A** matrix and $\underline{b}$ vector (see Equation A-4) into the linear software packages (e.g. LINGO, LINDO Systems, Inc. 2003). 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 A-1. All solutions met all the constraints without any difficulties and in a timely manner (e.g. less than a second per matching).

With the combination of the powerful personal computers and the zero-one linear programming, incorporating the component of multiple sets of matched samples into the statistical control modeling become very feasible. In most cases, such task (e.g., 1000 sets of matched samples along with 1000 times of ANCOVA analyses) can be completed during night time when researchers leave the office.

2. Statistical Modeling Followed by the Matching Method

After the matching procedure, a small pretest score (or other continuous covariates) difference between the treatment group and its matched sample may remain. The ANCOVA can be used to control for the effects of the small differences on covariates between the two groups. When the matching control is integrated with the ANCOVA analysis, ANCOVA results in *adjusted posttest means* for both groups under the constraint of two groups' pretest (or other continuous covariates) means as well as two groups' demographic backgrounds being equal. Those constraints make the ANCOVA-based adjusted means more defensible because no extreme extrapolation and/or a regression artifact are applied on the estimates of the adjusted means for both groups.

3. The Feasibility of Successfully Creating Matched Samples

The relative success in creating a set of matched samples relies on which covariates to base the matching. In general, when more important covariate variables are used in the process of matching, the result of the matched group is more similar to the experimental group.

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

4. Caveats

As with other matching methods, the matching method introduced in this study requires as many important covariates as possible. It also requires a relatively large non-treated population. When those conditions are met, a more reliable result is expected to be obtained, compared with any results produced by the evaluation design with only a set of matched sample used as a control group.

Nevertheless, causality cannot be inferred from the studies that incorporate the proposed matched method since the data are collected from quasi-experimental designs. The only designs that allow for relatively unambiguous causal inference are true experiments. The strength of a true experiment is its ability to rule out threats to internal validity through random assignment of students to treatment and control groups. The procedure of random assignment ensures that the study groups are equated on all possible observed and hidden variables. This is particularly true for the cases with large sample sizes.

References

Cohen, J. (1988). *Statistical power analysis for the behavior sciences*, (2nd Ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.

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.

Li, Y. H., Yang, Y.N. & Modarresi, S. (2005, April). Utilizing the zero-one linear programming constraint to draw a set of matched samples from a non-treatment population as a control group for the quasi-experimental designs. Paper presented at the annual meeting of the American Educational Research Association (AERA 2005), Montreal, Canada.

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.

Raudenbush, S. W. & Bryk, A. S.(2002). Hierarchical linear models: Applications and Data Analysis Methods, 2nd ed. Newbury Park, CA: Sage Publications, Inc.

Rosenbaum, R. R. (1998). Multivariate matching methods. Encyclopedia of Statistical Sciences, 2, 435-438.

Rosenbaum, R. R. (2002). Observational studies, 2nd ed. New York: Springer-Verlag.

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.& Thomas, N. (1992a). A finely invariant matching methods with ellipsoidal distribution. The Annuals of Statistics, 20, 1079-93.

Rubin, D. B.& Thomas, N. (1992b). Characterizing the effect of matching using linear propensity score methods with normal covariates. Biometrika, 79, 797-809.

Rubin, D. B.& Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. Biometrics 52: 249-64.

Rubin, D. B.& Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association, 95, 573-585.

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. (2005, April). Using the multiple-matched-set-of-matched-sample and statistical controls to examine the effects of magnet school programs on the reading and mathematics performance of students. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Appendix A:
The Updated Version of Li et al.'s (2005) Matching Method

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, etc.). 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). This Appendix provides a description of Z*ero-One Linear Programming* followed by detailed steps of using this technique in creating a set of matched samples as a control group.

A. Zero-One Linear Programming

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

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

Such that

$$A_{11}x_1 + A_{12}x_2 + \ldots + A_{1n}x_n \square b_1 \qquad \text{(A-2)}$$
$$A_{21}x_1 + A_{22}x_2 + \ldots + A_{2n}x_n \square b_2$$
$$\vdots \qquad \ldots \qquad \vdots$$
$$A_{m1}x_1 + A_{m2}x_2 + \ldots + A_{mn}x_n \square b_m$$

$$x_i \in \{1,0\} \qquad \text{(A-3)}$$

where

ABS is the absolute function,

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

P_i is the pretest score for member i without receiving a specific treatment,

P_j is the pretest score for member j in the quasi-experimental group,

A_{mxn} is the coefficient, and b_m is the right-hand side value for the m^{th} constraint (refer to LINDO API, p1, 2003),

 is the relationship function, which could be ≤, =, or ≥. The equal symbol of '=' is used for the category-constraint variables. The symbol of '≤' or '≥' is used for the continuous-constraint variables.

More specifically, in Equation A-1, the members in the non-treated 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 A-3 to identify whether the members are selected or not for the matched group. For example, if the number of students in the treatment and non-treatment groups are 3, 7 respectively, and we need to draw 3 students as a set of matched samples from the 7 non-treatment students, if the solution is $\underline{x}$'=[1 0 0 1 0 0 1], it means that the first, fourth and last member in the non-treatment groups will be selected as a set of matched samples. In the other words, the first position of "1" in the '$\underline{x}$' standards for the first member in the non-treatment groups will be selected; the second position of "0" in the $\underline{x}$' standards for the second member in the non-treatment groups will be not selected.

Equation A-2 introduced above can be presented by a matrix expression—Equation A-4 (shown below) in which the vector of $\underline{x}$ will be resolved by not only maximizing (or minimizing) the linear function of Equation A-1, but also imposing the constraint of x values of either one or zero. The matrix **A,** representing the left side of Equation A-2, and the vector $\underline{b}$, representing the right side of Equation A-2. Equation A-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 solve (refer to LINDO API, p1, 2003, Theunissen, 1985, 1986;van der Linden & Boekkooi-Timminga, 1989).

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

Several examples of the setting constraints for matrix **A** and the vector b can be: (1). The total days students attended at school for the set of the selected samples should be close to the ones in the quasi-experimental group, (2). The total family income for the set of the selected samples should be close to the ones in the quasi-experimental group, (3). The set of matched samples should have 5 type-I students as the ones in the quasi-experimental group, (4). The set of matched samples should have 8 type-II students as the ones in the quasi-experimental group, etc. The definition of types of students has been explained previously.

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 A-1. The choice of types of composite score can depend on the nature of those pretest scores. The following section will introduce a matching method utilizing the zero-one linear programming.

B. The Greedy Combined with the Optimal Matching Method

As indicated in the summative evaluation described in this paper, there were 155 and 1505 students in a reading intervention and non-reading intervention programs, respectively. The number of treated sample and non-treated population will be used as an example in this section. Since the matching method introduced by Li et al.' (2005) weighted the pretest-score variable more than the rest of covariate variables, the pretest-score variable was treated separately from the other covariates.

In Equation A-1, if the pretest score of P_j is set to the mean of the quasi-experimental group and the vector of x is resolved, we are able to seek a set of matched samples (e.g., 155) from non-treatment population (e.g., 1505). This set of matched sample has exactly one-by-one category-type demographic (e.g., gender, race, etc.) variables as the ones in the treatment group. In addition, this set of matched samples has very similar summary statistics on continuous (e.g., total days that students attended schools) variables and the pretest variable as the ones in the treatment sample.

Nevertheless, the "pretest-mean matching" approach introduced above is not desirable because the pretest scores for each selected matched sample tend to be close to the pretest mean scores of the treatment group. This causes the variability of the pretest scores for these matched samples to be too small and not as similar as its counterparts in the quasi-

experimental group. The proposed matching method is described in the following section. The details of the procedure are given in Figures A-1 throughA-3.

Instead of entering the pretest mean of quasi-experimental group into the targeted function of Equation A-1, the first member's pretest, P_j in the quasi-experimental group is entered into the targeted function. Once the vector of $\underline{x}$ is resolved, we are able to identify 155 matched samples from 1505 non-treatment population that has identical category-type (e.g., gender, race, etc.) covariates and similar interval covariates when compared to the treatment group. Afterwards, a member is selected among 155 matched samples with the following conditions, a) Has the same type of member as this individual member, and b) Has the closest pretest score to this individual member. Once this member is found, he/she will be the matched member of the first member of the quasi-experimental group (refer to Figure A-1 shown in the Appendix A).

For selecting the matched sample from the non-treatment population for the second member of the quasi-experimental group, the researcher should replace any members that have not been previously selected in the non-treatment population pool. Next, he/she should 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 is included in the next set of samples. In addition, the researcher should enter the second member's pretest score in the targeted function. Finally, when another new set of 155 matched samples is drawn form 1504 non-treatment population, the researcher should find a member from this new set of 155 samples with the following conditions, a) Has the same type of member as this individual member, and b) Has the closest pretest score to this individual member. Once this member is found, he/she will be the matched member of the second member in the quasi-experimental group (refer to Figure A-2 shown in the Appendix A).

For selecting the matched sample from the non-treatment population for the third member of the quasi-experimental group, the same steps taken for the second member should be repeated. That is, replace any members that have not been previously selected in the non-treatment population pool. Then, add an additional constraint to the constraints that have been imposed in the *zero/one* linear model to ensure that all of the members being recently selected are included in the next new-drawn set of samples. In addition, the third-member's

pretest scores should be entered in the targeted function (refer to Figure A-3 shown in the Appendix A).

Repeat the steps taken for the third member until all members in the quasi-experimental group have found their own respective members from the non-treatment population. The above individually-based matching procedure begins with drawing a set of samples that meets all the desirable constraints and has smallest distance to the member's pretest score as possible given the condition that the previously selected matched sample is included in this set of samples. A member who meets the required criteria stated above is then selected from this set of samples, instead of directly from the full non-treatment population. Each of the next serial set of samples includes all previously selected members and consequently the set of matched samples that researchers desire that meet all constraints. It is important to note that sampling without replacement is applied for creating a set of matched samples.

C. Creating Multiple Sets of Matched Samples: Matching Procedure Incorporating Measurement Error

A set of matched samples will be generated taking the steps above if we assume the pretest scores of a treatment group (e.g., reading intervention program students) is a true score, not contaminated with any measurement error. For large sample sizes, this assumption is appropriate. However, to increase the confidence level of seeking an appropriate matched sample as similar to the treatment group as it can be, researchers may allow the pretest-score to be contaminated with a "reasonable" measurement error. Equation A-5 helps us comprehend this concept.

$$\text{Minimize } \sum_{i=1}^{n}\left[ABS(P_i - (P_j + E_j))\right] x_i \qquad \text{(A-5)}$$

The components in Equation A-5 are the same as those found in Equation A-1, with the addition of measurement error component, E_j. The value of E_j 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,

$$SE^2 = \frac{S^2}{N} \qquad \text{(A-6)}$$

Where

N is the sample size of the treatment group,

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

By allowing measurement errors into the matching procedures illustrated in the proposed matching method (refer to Figures A-1 through A-3), a set of matched samples can be generated given a different value of measurement error to each member in the quasi-experimental group. After sampling, every member of the non-treatment group is returned to the dataset. Another set of matched samples will be generated given a different value of measurement error to each member in the quasi-experimental group. Again, every member of the non-treatment group should be returned to the dataset after sampling. After repeating the matching procedure again and again, multiple sets of matched samples will be created. It should be noted that many members from the population of the non-treatment group may appear multiple times in different sets of matched samples because the same constraints have been repeatedly imposed into the matching procedures.

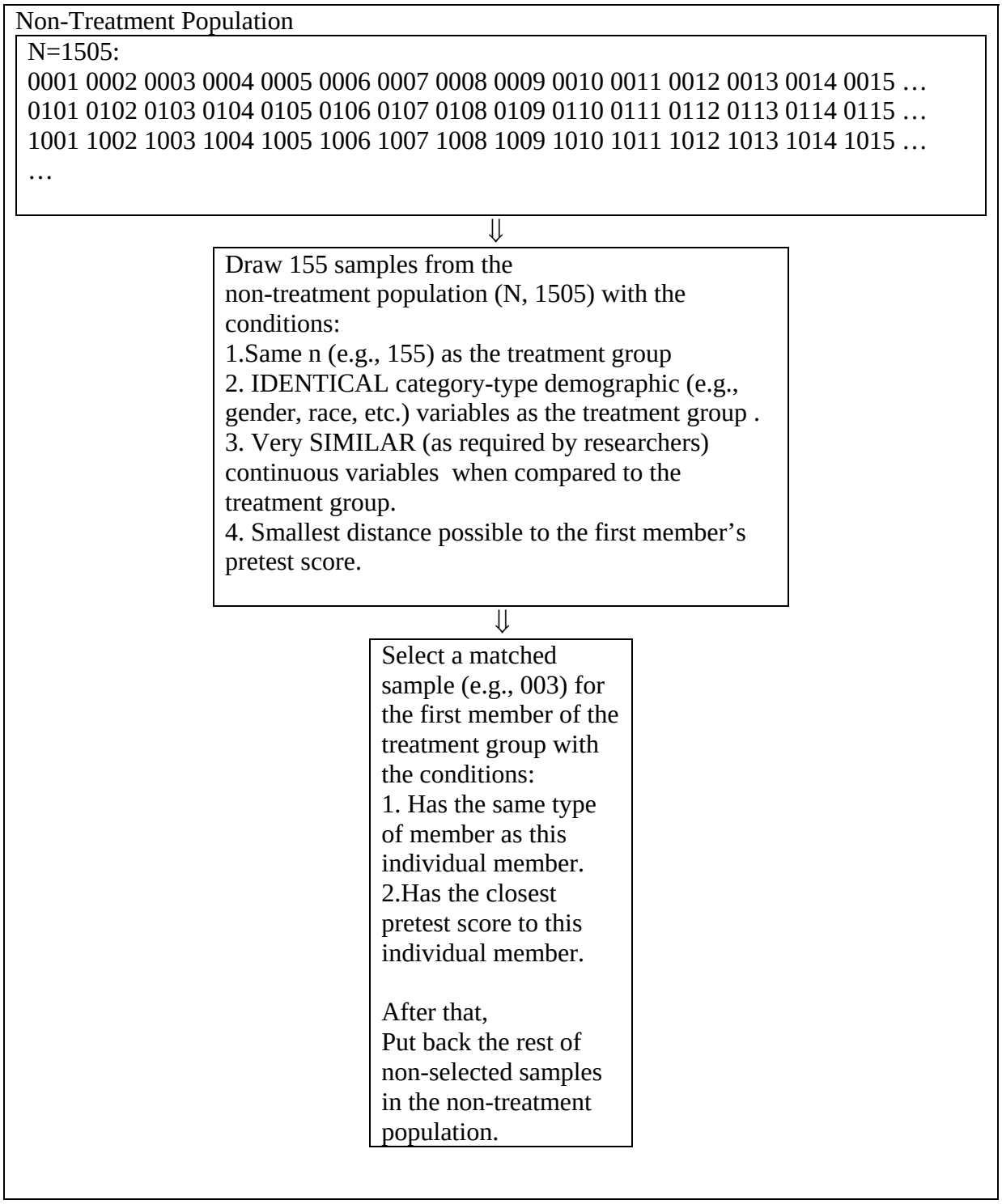


Figure A-1. The Procedure for Finding the Matched Sample for the First Member of the Treatment Group

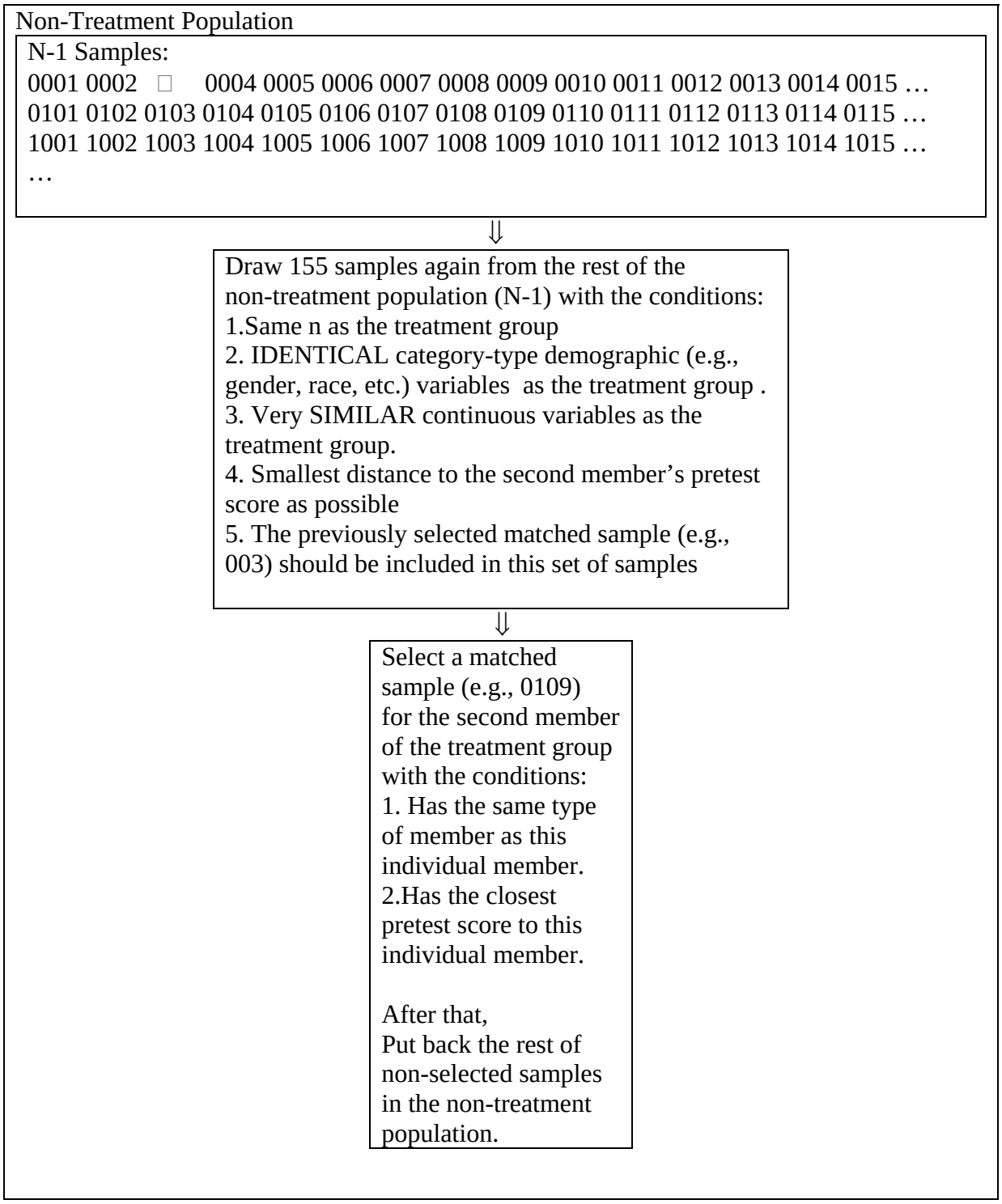


Figure A-2. The Procedure for Finding the Matched Sample for the Second Member of the Treatment Group

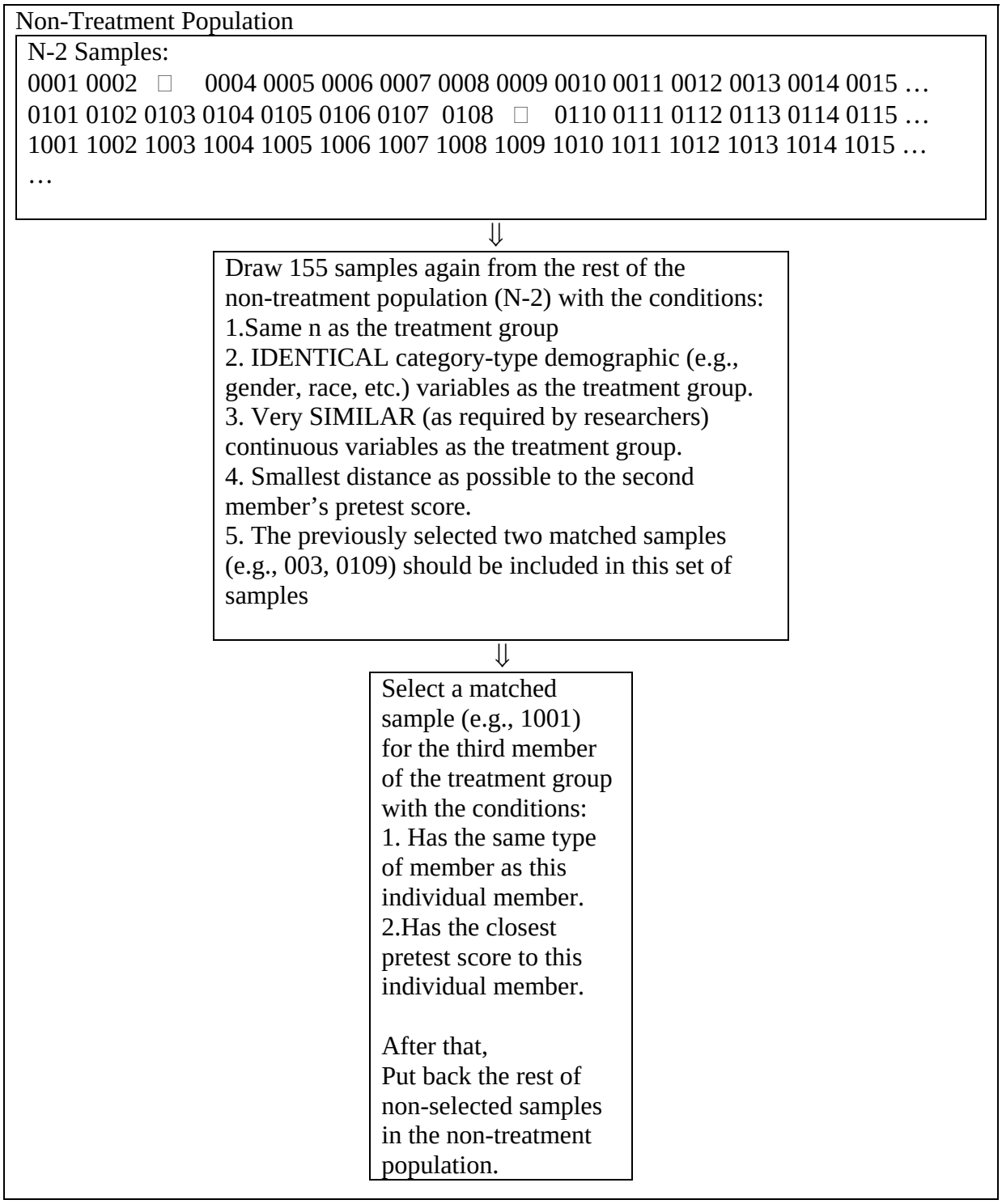


Figure A-3. The Procedure for Finding the Matched Sample for the Third Member of the Quasi-Experimental Group

Appendix B:
A Set of Matched Samples Created by the Greedy Matching Method with the Propensity Score

Table B-1:

Thirty Matched Pairs on Several Matching Variables (e.g. the pretest score, days of students attending at the school, sex, race, poverty status, attending school and the types of student) Using Greedy Matching Method with the Propensity Score

Paired	Propen		Type		Race		Gender		Poverty		School		Pretest		Days in Schools	
	Pro	Mat	Pro	Mat	Pro	Mat	Pro	Mat	Pro	Mat	Pro	Mat	Pro	Mat	Pro	Mat
1	.20	.20	3	36	3	3	1	2	1	2	1	3	354	374	171	173
2	.21	.21	3	1	3	1	1	1	1	1	1	1	352	397	175	163
3	.11	.11	3	36	3	3	1	1	1	3	1	4	379	344	147	150
4	.13	.13	3	5	3	5	1	1	1	1	1	1	365	342	148	162
5	.07	.07	3	36	3	3	1	1	1	3	1	4	405	390	146	178
6	.11	.11	3	30	3	3	1	2	1	1	1	5	386	410	162	171
7	.10	.10	3	36	3	3	1	2	1	3	1	2	390	390	152	172
8	.09	.09	3	18	3	3	1	1	1	2	1	3	390	410	147	155
9	.15	.15	3	15	3	3	1	1	1	1	1	5	363	367	157	143
10	.15	.15	3	15	3	3	1	1	1	1	1	5	351	376	137	153
11	.12	.12	3	6	3	3	1	1	1	1	1	2	379	360	157	152
12	.06	.06	5	15	5	5	1	1	1	3	1	1	401	394	165	159
13	.07	.07	5	30	5	5	1	2	1	1	1	3	383	397	154	172
14	.19	.19	6	12	3	3	2	2	1	1	1	2	362	347	171	172
15	.18	.18	6	18	3	3	2	2	1	3	1	1	360	363	166	174
16	.17	.17	6	18	3	3	2	2	1	1	1	3	356	362	155	167
17	.21	.21	6	3	3	3	2	1	1	1	1	1	342	348	156	171
18	.17	.17	6	3	3	3	2	1	1	1	1	1	367	351	166	149
19	.16	.16	6	30	3	3	2	2	1	1	1	5	362	388	154	172
20	.11	.11	6	30	3	5	1	1	1	2	2	3	383	377	178	173
21	.11	.11	6	12	3	3	1	2	1	1	2	2	383	383	177	171
22	.11	.11	6	9	3	3	1	1	1	3	2	1	381	394	170	170
23	.13	.13	6	27	3	3	1	1	1	3	2	3	362	383	160	175
24	.19	.19	9	9	3	3	1	1	3	3	1	1	356	354	174	170
25	.14	.14	9	9	3	3	1	1	3	3	1	1	377	379	175	178
26	.13	.13	9	18	3	3	1	2	3	3	1	1	370	370	158	152
27	.08	.08	9	5	3	5	1	1	3	1	1	1	406	367	161	148
28	.12	.12	9	12	3	3	1	1	3	2	1	2	374	395	153	178
29	.12	.12	9	48	3	3	1	2	3	2	1	4	384	367	169	160
30	.12	.12	9	30	3	3	1	2	3	1	1	5	381	409	163	174

Table B-2.
Descriptive Statistics of the Pretest Scale Score and the Total Days Students Attended at Schools for the Reading Intervention and Matched Sample Groups Using Greedy Matching Method with the Propensity Score

Group	N	Type of Students	Pretest Scale Score				Days of Student Attended at Schools			
			Mean	SD	Min	Max	Mean	SD	Min	Max
Treatment	155	Different	371	19	311	418	167	10	129	178
Matched	155		375	23	317	441	167	9	132	178

Appendix C: The Interpretation of ES

The standardized ES can be thought of as the percentile rank (PR) standing of the treatment sample mean when compared with the distribution of the multiple-matched-sample test scores (Cohen, 1988). For example, if a particular program treatment has 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 principle developed by Cohen (1988), the look-up table presented below is used to interpret the meaning of the ES in terms of its PR standing in the match sample. (See Table 5)

Table C-1
Converted Effect Size (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