

*DRAFT Report– Please do not cite or distribute without permission*

*DRAFT - Please do not cite or distribute without permission of author*

**Experimental and Nonexperimental Estimates of Program Impact  
Using the Tennessee STAR Experiment**

Russell Cole

University of Pennsylvania

Paper Presented at the Annual Conference  
of the American Education Research Association

March 28, 2008

New York, NY

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C050041-05 to the University of Pennsylvania. The opinions expressed are those of the author and do not represent views of the U.S. Department of Education.

## **Abstract**

This paper assesses whether a nonequivalent comparison group can be used as a replacement for a randomized control group in the estimation of the causal impact of the Tennessee STAR experiment when a limited set of covariates are available for equalizing the groups. Two sets of program impacts are calculated—one using the treatment and control groups from the random experiment, and one using the treatment group from the experiment and a nonequivalent control group. Propensity score adjustment is used to try to reduce the bias in the estimation of the causal effect in the hypothetical observational study. Sensitivity analyses were performed to assess the adequacy of the propensity score in creating equivalent groups according to the known variables of interest. In this paper, the limited set of covariates were not sufficient to create equivalent groups, and the program impacts calculated from the propensity score matched pairs were markedly different from the “true” program impact estimates. It was also noted that conservative propensity score matching procedures and the matching of students in the treatment condition to the control condition without replacement were the methods that produced fewer discrepancies between experimental and nonexperimental impact estimates. The final recommendations of this paper serve as a warning for researchers interested in making casual statements from limited observational data.

## **Introduction**

Estimating the effect of a program requires the comparison of individuals who have received the service of the program (the treatment condition) against individuals who have not received the service of the program. In order for the comparison to be both fair and valid, the untreated comparison group must not differ systematically from the treatment group on any variables that are related to the outcome of interest. The only way to ensure that all variables (measured and unmeasured) are equally distributed across the treatment and comparison group is through random assignment of the treatment condition to individuals in the study (Boruch, 1977).

In practice, estimating the effects of a program through a randomized controlled trial (RCT) is often expensive and difficult to implement. Program stakeholders may wish to rely on cheaper alternatives for measuring program impact. A number of important studies have attempted to compare the effectiveness of estimating program effectiveness through nonexperimental methods as opposed to through experiments (Fraker & Maynard, 1987; LaLonde, 1986; Friedlander & Robins, 1995; Dehejia & Wahba, 1999; Wilde & Hollister, 2002; Agodini & Dynarski, 2004). These studies have met with mixed results: In some cases propensity score matching was demonstrated to provide program impact estimates that were similar to the estimates obtained from a randomized experiment (Dehejia & Wahba, 1998, 1999).

This paper will follow in the tradition of papers that compare “true” program impact estimates obtained through randomized experiments with the program impact estimates obtained through the creation of a comparison group using propensity scores. Rosenbaum and Rubin (1983) have demonstrated that the use of propensity scores can be used to select a comparison group that is similar to the treatment group according to all measured characteristics. In this

paper, I will test the feasibility of using propensity scores to create a cohort for comparison when comparison data are limited, and I will illustrate the different program impact estimates obtained from an experimental and a (created) observational study.

Borrowing from the framework of Agodini and Dynarski (2004), this paper will address the following questions:

- How feasible are propensity score methods when limited data are available for constructing comparison groups?
- How well do propensity score methods replicate experimental impacts associated with small class membership on four student achievement outcomes—Reading, Math, Listening, and Word Skills?

For the purposes of this paper, I will use data from the Tennessee Project STAR experiment, an undertaking that has been described as “one of the great experiments in education in U.S. history” (Mosteller, Light, & Sachs, 1996, p. 814). The Student/Teacher Achievement Ratio (STAR) project sought to determine the effects of reduced class size on elementary (K-3) students’ academic achievement. The effects associated with small class size have been reviewed by Glass and Smith (1979), Glass, Cahen, Smith, and Filby (1982), Hedges and Stock (1983), Mosteller, Light, and Sachs (1996), and Hanushek (1999). With the exception of the Hanushek piece, the syntheses of research suggest that increasingly positive effects are noted as classes become smaller (Nye, Hedges, and Konstanopolous, 1999). As the purpose of this paper is to compare program impact estimates and not to revisit the literature on the effects of class size, a brief description of the program/data is illustrated below.

### **Background on Project STAR**

The STAR project was a randomized experiment that was initially commissioned in 1985 by the Tennessee state legislature. During the 1985-86 school year, kindergarten students were randomly assigned to one of three conditions: small class size, regular class size, or regular class

size with a full-time teacher aide. Classes were considered small if they contained between 13 and 17 students, whereas regular sized classes had 22-26 students. (Zaharias, Achilles, & Cain, 1995). Students who entered the study at later times were randomly assigned to classes upon entry.

Seventy-nine schools in 42 school systems participated in the STAR experiment. The design of the STAR experiment involved the random assignment of students to teachers and of teachers to treatment conditions within schools. Due to the randomization of students and teachers into treatment and control conditions, the program impact estimates are calculated by comparing treatment and control classes *within* each school. In the estimation of the STAR program impact, Krueger (1999) found that, on average, students in small classes scored significantly higher on standardized tests than students in regular sized classrooms. These impact estimates using the experimental data are the best possible estimates of the “true” program impact, and are appropriate considering the research design.

Due to the burdensome nature of this design, there were a number of schools that chose to not participate in the random assignment procedure, but did agree to have achievement and demographic data collected on all of their students. The STAR database contains data from 11,601 students involved in the randomized experiment, as well as data from 1,780 students from 21 comparison schools not participating in the STAR experiment. STAR principal investigators believed that these 21 comparison schools were similar in terms of operational characteristics to the existing 79 STAR schools, and as a result, these comparison data might serve as an adequate nonequivalent group from which to try to estimate the causal effect associated with small class size.

### **Previous propensity score work using Project STAR data**

Wilde and Hollister (2002) explored nonexperimental estimates of impact against experimental estimates using STAR data. In their study, they capitalized on the within-school randomization design to compare students who received the treatment in one school with similar students in the control groups in different schools. As indicated before, the “true” program impact estimate is calculated by comparing students *within* a school who received the treatment as opposed to the control condition. Wilde and Hollister used propensity score matching to create the nonexperimental comparison group across schools, and then compared the program impact estimate generated from the experimental and observational versions of the study. These researchers concluded that the nonexperimental estimates were not very “close” to the “true” experimental estimate of program impact.

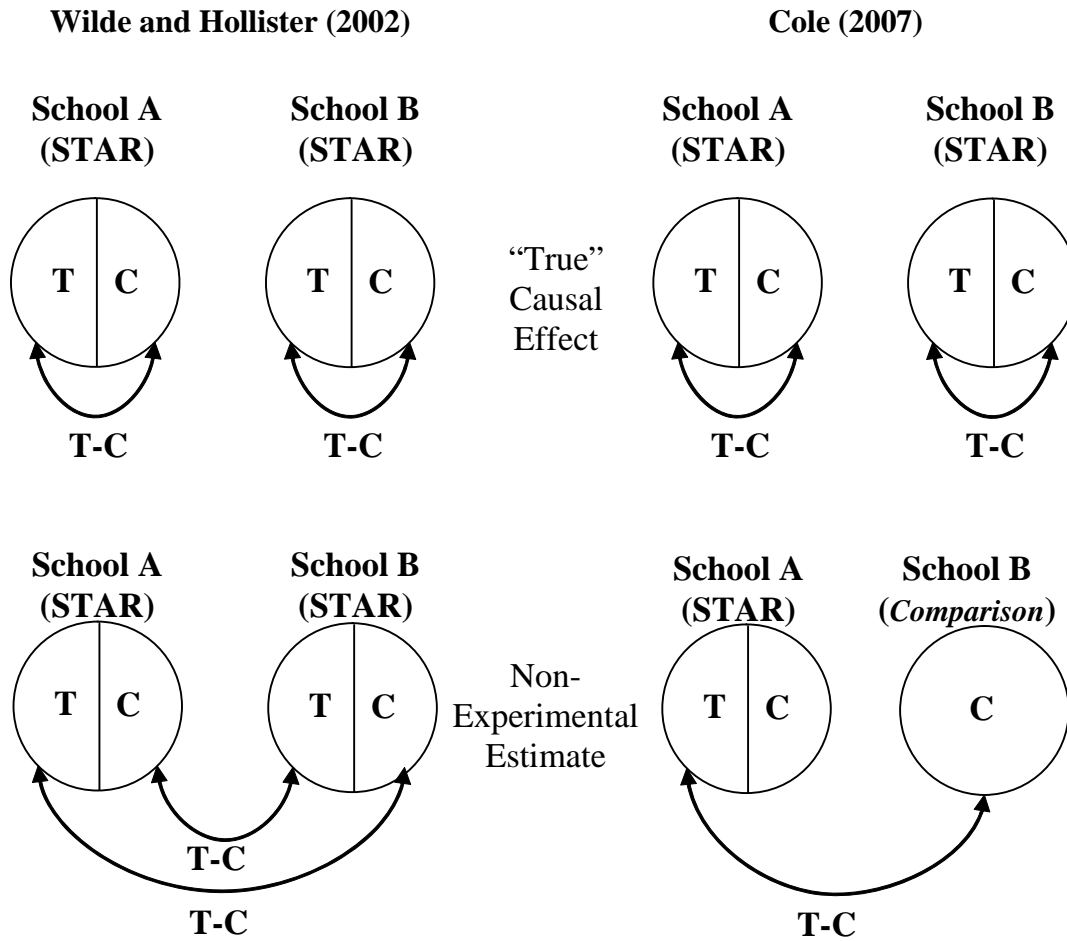
One of the major benefits to the Wilde and Hollister design was that they were able to take advantage of the host of demographic variables that were collected in the STAR project in the creation of their propensity scores. Above and beyond typical demographic variables (age, sex, race), the investigators of the STAR program collected data on student free-lunch status, teacher race, teacher education, teacher experience, etc. Due to the fact that Wilde and Hollister used data from within the STAR experiment, they were able to exploit these and other demographic variables in their estimation of the probability that each student was a member of the treatment condition. However, in a more realistic setting, the available data from a nonequivalent comparison group would not perfectly mirror the data collected for the individuals receiving the treatment. As such, the exercise reported by Wilde and Hollister may be idealistic in terms of the creation of the propensity scores for subjects and, therefore, this topic merits further attention with the Tennessee STAR data.

### **Analysis Framework**

As stated previously, this paper will compare the program impact estimates associated with experimental and nonexperimental data from Project STAR. Unlike Wilde and Hollister (2002), I will use the 21 comparison schools that were not involved in the STAR program as the source of control subjects for my comparison group. Figure 1 indicates the different method whereby the nonexperimental program impact estimate will be obtained in the current study.

Both this paper and the Wilde and Hollister paper estimated the “true” causal impact within each school. In Wilde and Hollister, the nonexperimental program impact estimates were captured by comparing students in the treatment condition in one school to students in the control condition in a different school. In this current study, the nonexperimental impact estimate will be calculated by comparing the students in the treatment condition in the STAR experiment to a set of students in schools not participating in the STAR experiment.

**Figure 1: Program Impact Estimate Logic**



For the purposes of this paper, I will concentrate solely on two of the treatment conditions for measuring program impact: one, students who were members of a small class (T) and two, students who were in regular class (C). In the comparison sites, it is assumed that the students were all in “regular” classes in terms of size and that there were not classroom aides in these sites. As such, the only comparisons that are viable in the experimental and nonexperimental settings are between students who were in a small class versus students who were in a large class.

Unlike Wilde and Hollister (2002), very few demographic variables were collected in the comparison sites. Only student age, gender, and race were available in this dataset. This much

smaller subset of variables available for propensity score matching severely limits the likelihood that students will be matched on all unmeasured variables related to the outcomes of interest, but does provide a more realistic example of the use of existing administrative data as a potential source for the creation of a comparison cohort.

In addition, student data for the comparison sites only contains outcome data for grades 1-3. The STAR random assignment is intended to take place during kindergarten, and as such, program impact estimates should begin in this year of schooling. However, due to the lack of existing data, program impact estimates in the nonequivalent comparison group are only tenable in grades 1-3, and not K-3 as in the experimental analysis of the STAR study.

### **Data**

The data for this paper were obtained from HEROS (Health and Education Research Operative Services) at <http://www.heros-inc.org/>. Individual records for students in both the STAR experiment and the comparison schools are publicly available at this website.

As indicated above, demographic data (age in days, race, and sex) were available for each student in both the STAR experiment and in the nonequivalent comparison group for grades 1-3. In addition, school and teacher ID variables were included in each record, along with the scale score for four SAT tests: Reading, Math, Listening, and Word Skills.

Students were considered for this paper only if all data for demographics and achievement tests for a given grade were available, and none was missing. Table 1 in the Appendix contains descriptive statistics for all observations and all variables used in the current study. It is worth noting that approximately 2-3 times more observations exist for the STAR treatment group than the nonequivalent comparison group, which is not surprising given the number of schools involved in the STAR experiment. It also is of interest that the nonequivalent

comparison group had far fewer minority students and it appears that their test scores were generally higher than both the STAR treatment and control groups on the achievement outcomes in most grades.

### **Methods**

In order to estimate the experimental impact of small class sizes, the following regression equation (borrowed from Krueger, 1999) was employed for students in each grade level:

$$(1) \quad Y_{ics} = \beta_0 + \beta_1 SMALL_{cs} + \beta_3 X_{ics} + \alpha_s + \mu_{cs} + \varepsilon_{ics}$$

Where

- $Y_{ics}$  is the scale score on the given SAT test of student  $i$  in class  $c$  in school  $s$
- $SMALL_{cs}$  is an indicator variable for whether the student was assigned to a small class in that grade
- $X_{ics}$  is a vector of student covariates (age, gender, race)
- $\alpha_s$  is a fixed effect (dummy variable) for each school, which allows for the within-school experimental impact to be unbiased
- $\mu_{cs}$  is a random effect for each classroom  $c$  in each school  $s$
- $\varepsilon_{ics}$  is the error term for student  $i$  in class  $c$  in school  $s$

The parameter  $\beta_1$  indicates the small class experimental impact estimate for a given test in a given grade, and will be the variable of interest for comparison with the nonexperimental program impact estimation.

To create the nonexperimental impact estimate, propensity score matching was employed to create pairs of similar students across the available demographic variables. In each pair, there was one student who received the small class treatment condition and one “similar” student selected from the 21 nonequivalent comparison schools. In order to estimate the propensity scores for each student in each grade level based on the observed data, logistic regression was used. The dependent variable in this regression was whether or not the student was in the

treatment condition; the independent variables were each student’s demographic characteristics. As noted earlier, the data for this estimation were obtained by taking *only* students in the STAR experiment who received the treatment (T cases) and all of the students in the 21 nonequivalent comparison sites as the controls (C cases). The following logistic regression was employed:

$$(2) \quad \log \left[ \frac{\Pr(Z_i = 1 | \mathbf{X}_i = \mathbf{x}_i)}{1 - \Pr(Z_i = 1 | \mathbf{X}_i = \mathbf{x}_i)} \right] = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Where

- $Z_i$  is the indicator of whether or not student  $i$  received the treatment condition ( $Z_i=1$  indicates that student  $i$  was in a small class in the STAR experiment,  $Z_i=0$  if  $i$  was a student in the comparison schools)
- $X_i$  is a vector of student covariates including age, gender and race (as well as interactions between all variables as well as higher order terms to increase the predictive accuracy of the regression given the paucity of covariates)
- $\varepsilon_i$  is the error term for student  $i$

Estimated propensity scores were calculated for each individual where each propensity score indicated the probability of being a member of the treatment group given one’s own characteristics. In order to check the accuracy of the propensity score method in creating equivalent groups, a sensitivity check was employed based on the stratified propensity scores. The estimated propensity scores for each individual were sorted and divided into bins of approximately 150-175 students (Agodini & Dynarski, 2004). Observations in the treatment group with higher propensity scores than the highest propensity score in the control group were eliminated, and observations in the control group with lower propensity scores than the lowest propensity score in the treatment group were eliminated due to a lack of common support (Dehejia & Wahba, 1999). Within each bin, average propensity scores were tested between the treatment group and the control group to determine if the groups were equivalent within strata.

Next, each of the covariates in the model was compared across the treatment and control groups within each stratum to see if these variables were indeed balanced. If there was a statistically significant difference between the groups on a given covariate, interactions and higher order terms were added to the model until balance across all covariates was achieved.

A “greedy” matching algorithm was employed to select individuals from the comparison group that were as similar as possible to members of the treatment group according to the propensity scores (i.e., the *nearest neighbor*). This algorithm first linked individuals who were the most similar (those whose propensity scores were nearly exactly equal) and progressively matched less and less similar students until the matches were either too poor (i.e. scores were  $> .1$  unit apart) or until all the students in the treatment condition were matched. For one set of impact estimates, members of the comparison group were selected with replacement, since there were fewer members in the comparison group than the treatment group, and since the two groups appear to be different in general (see Table 1). As such, a potential comparison group member could be matched to more than one treatment group member (Agodini & Dynarski, 2004). Impact estimates from the propensity matched pairs were also calculated for matching without replacement. In sum, there were two constructed nonequivalent comparison groups used for this study—matching with and matching without replacement.

When individuals are matched to each other, the researcher must decide whether to optimize the number of matches or the quality of the matches. It is possible to create a large number of potentially poor matches, or a relatively sparse number of excellent matches. In this paper, I will investigate the decisions that result with varying specifications of the quality of matches according to the propensity score. I will compare the results that are obtained when individuals match on varying digits on their propensity scores. The most conservative matching

procedure will match individuals only if the first 5 digits of their propensity scores are identical (i.e., their propensity scores differ by less than .00001). Less conservative matches (4, 3, 2, 1 digit matches) will also be considered, and program impact estimates will be calculated for each of these possible decision rules. The matching algorithm is operationalized using the “greedy” macro for SAS (Parsons, 2001).

The matching of treatment to control members was selected as the method of computing the treatment effect over alternate methods of assessing program impact with propensity scores (i.e., use of propensity score as a covariate, stratification based on the propensity score) due to the recent research of Victor (2007). In this unpublished dissertation, it was demonstrated that pair-wise propensity score matching had the smallest overall bias in the estimation of effects in an observational study as compared with other estimators (Mahalanobis Matching, Cluster Analysis, etc.). In the current study, the difference in outcomes across two matched individuals (one who received the small class treatment and one nonequivalent comparison student) is therefore the method used to establish the causal impact of small class sizes.

Program impact estimates from the propensity score matched pairs are estimated according to the following equation (for each subject in each grade):

$$(3) \quad Y_{ip} = \beta_0 + \beta_1 SMALL + \mu_p + \varepsilon_{ip}$$

Where

- $Y_{ip}$  is the scale score on the given SAT test of student  $i$  in pair  $p$
- $SMALL$  is an indicator variable for whether the student was assigned to a small class in that grade (i.e., whether or not the student was in the STAR program or the comparison school)
- $\mu_p$  is a random effect for each matched pair  $p$
- $\varepsilon_{ics}$  is the error term for student  $i$  in pair  $p$

The individual covariates (sex, race, age) originally used in the first equation are not necessary in this specification, due to the random effect for each pair  $p$ . Since pairs of students have relatively equal propensity scores, and since the propensity scores were demonstrated to equally divide covariates across treatment groups, this source of noise is eliminated. The parameter  $\beta_1$  indicates the small class nonexperimental impact estimate for a given test in a given grade, and will be the variable of interest for comparison with the experimental program impact estimation.

### **Results**

The experimental impact estimates associated with small class size are indicated in Table 2 in the Appendix. A (positive) statistically significant effect of small class size was noted in all tests and in all grades with the exception of Grade 3 Listening Skills. The estimated experimental impacts ranged from a low gain of approximately 1.25 points in Grade 3 Listening Skills to a high of approximately 14 points in Grade 1 Reading. Standardized effect sizes for these estimates ranged from .04 to .28 standard deviation units, with an average standardized effect size of .19 across all tests and all grades. Thus, these impact estimates demonstrated a positive effect on student achievement caused by being a member of a small class, and serve as the standard against which the nonexperimental impact estimates can be compared.

The validity of the results from the nonexperimental impact estimate is conditional on the adequacy of the propensity score matching performance. The tests of the propensity scores across the treatment and nonequivalent comparison group strata are not presented, nor are the

tests of each covariate within each stratum for the sake of brevity<sup>1</sup>. Instead, the findings are described in brief below.

In the estimation of propensity scores for Grade 1, the given set of covariates was not capable of creating equivalent strata of propensity scores or matched covariates. Despite the inclusion of a number of interactions and higher order terms, I was unable to create an equation that balanced both propensity scores and covariates across the groups.<sup>2</sup> The final solution obtained for Grade 1 contained one marginally significant difference ( $p = .04$ ) across one of the strata with respect to differences in propensity scores, and one covariate (log of age in days) was marginally significantly different across the two groups within this stratum ( $p = .05$ ). As this was the best possible solution, it was used for the calculation of program impact estimates, with one caveat—assumptions of equivalence across groups have been clearly violated.

Equivalence across the groups was obtained in the Grade 2 sample. Using approximately 150-160 students per stratum, propensity scores and all covariates were balanced between the treatment and control group. There were no statistically significant differences found between these groups on any measured covariates or in the calculated propensity scores within each stratum.

Equivalence across the groups was nearly obtained using the Grade 3 sample. Approximately 155 students were used per stratum. All covariates were balanced within each stratum (there were no significant differences in covariates at the  $p < .05$  level). However, one of the stratum had a statistically significant difference in mean propensity scores ( $p = .04$ ). As with

---

<sup>1</sup> The final logistic regression specifications were different across the three grades of estimations. In all grades, main effects for age (in days), gender, and race were included. Interactions between these variables, as well as higher order terms and log transformations of the age variable (along with necessary interactions), were introduced one at a time to help balance covariates across groups.

<sup>2</sup> Different strata sizes were also considered, but no combination of interaction effects and different bin sizes were able to fix the significant differences in either covariates or propensity scores across treatment conditions.

the Grade 1 results, this was the optimal solution obtained with these data. This solution was used for the program impact estimation, but once again it is not an optimal solution, and as such, the validity of the results is questionable.

The common support plots for the logistic regressions for each grade are presented as Figures 2-4 in the Appendix. Had the logistic regression performed well, we would expect to see a positively skewed distribution of propensity scores for the treatment group, with the mode of propensity scores close to 0. This would imply that the majority of students in the comparison group had a very low probability of being assigned to the treatment condition, based on their demographics. Similarly, with a well-performing logistic regression, we would expect to see a negatively skewed distribution for the treatment group, with the mode of propensity scores close to 1. Such a diagram would illustrate that the logistic regression had performed well at dichotomizing the students into their appropriate group, and that the sets of covariates used in the analysis were appropriate proxies for group membership.

In Figures 2-4, it is clear that the logistic regression has not performed particularly well in explicating group membership, in that the observed shapes of the graphs are markedly different from the expected shapes. The most alarming feature to these graphs is the clearly bimodal nature of the treatment group graphs. The comparison graphs appear to be relatively normal (perhaps slightly positively skewed, as expected). However, in spite of these issues, there was adequate overlap in the propensity scores (generally where propensity scores ranged from .22 to nearly 1.00). Given that the propensity scores had done a relatively decent job of distributing covariates across the treatment and control conditions, the last step was to create pairs of students, matched on the propensity score, and to calculate impact estimates.

Tables 3 and 4 present the impact estimates from propensity score matching calculated with and without replacement of the comparison group, respectively. The results of Table 3 will be described first. Of the 60 estimations of test scores in the three grades, 18 of the propensity score estimates provided statistically significant results that were appropriate (same direction and significance as the “true” program impact estimated from the randomized experiment). On seven occasions, the impact estimates from the propensity scores were negative and statistically significant, when the true program impact estimate was actually positive. As such, the propensity results demonstrated that small class sizes were significantly *worse* than regular classes in seven of the program impact estimations. The propensity score matched estimates correctly did not falsely reject the null hypothesis during the five estimations of the Grade3 listening scale score. It is important to note that in the estimation process for Grade 2— the one grade where the sensitivity check of the propensity matching estimates appeared valid— there were no significant estimates found for any of the test results under any of the five matching criteria.

There did not appear to be any systematic difference in the performance of the estimates under different matching decision criteria. When individuals were matched according to the most conservative decision rule (i.e., their propensity scores must not differ by more than .00001), the results were not consistently different than when individuals are matched according to the most generous criterion. There were two notable exceptions, however, in Grade 1. The results for the reading test were (correctly) positive and significant for the conservative approaches, and the estimates became nonsignificant as the more generous matching criteria were used. Also, the results for Math were nonsignificant under the most conservative criteria, but became statistically significant in the *wrong* direction under the more generous criteria. With the more conservative approaches, far fewer individuals were available to be used in the matching process.

Under the most conservative approach in Grade 1, only 588 (=1176/2) of the individuals in the treatment condition found a suitable match in the comparison group in this matching with replacement analysis. With the more generous criteria, all 1,607 of the students in the treatment condition found a match.

Table 4 indicates the results that were obtained when matching without replacement was considered. The first difference to note in these tables was the difference in the number of matched pairs that were obtained. In one case, only 34% of the pairs were maintained from matching with replacement to matching without replacement (Grade 3, Propensity score match < .1). On average, approximately 54% of the pairs were maintained when replacement was not allowed. This indicates that in the initial analysis (matching with replacement), there were a number of students in the comparison group who were “used” multiple times in the creation of pairs. This is not surprising, since there were more students in the treatment condition than in the comparison group. It is of interest to note the ways in which students were used multiple times: As the comparison group contained far fewer nonwhite students, these students played an increasingly larger role in the formation of matched pairs than their white counterparts. Thus, these students became key players in the estimation of the program impact, merely due to the fact that they represented a minority in the population.

The program impact estimates obtained from propensity matching without replacement were relatively similar to those obtained from matching with replacement. On 16 out of the 60 estimates, the propensity estimates were correctly positive and statistically significant. Only five of the program impact estimates were in the wrong direction and statistically significant (Grade 1 Word Study Skills). The 10 estimates associated with this test were negative and statistically significant in estimation with and without replacement in Grade 1, when the “true” program

impact was positive and significant. The differences observed may just be an artifact of the fact that the descriptive statistics in Table 1 indicate that the comparison group's word study skills scores were on average much higher than the average STAR small class results. Finally, similar to the matching with replacement results, the null hypothesis of no difference was correctly not rejected for the five impact estimates of Grade 3 Listening skills.

### **Discussion**

There are two stated goals of this research paper, both of which merit discussion. The first research question regards the feasibility of propensity score methods for estimating program impact when a limited set of covariates is available. In this empirical investigation, it was determined that while propensity scores can be calculated for individuals, that these propensity scores did not perform particularly well in explicating group membership, nor did these propensity scores create balanced strata across the treatment and control groups with respect to the covariates used in the estimation process. While Wilde and Hollister (2002) were able to capitalize on the within-school randomization to best use all of the available data for their propensity score matching, this paper may serve as a more realistic example of how propensity scores might be used with an existing data source. It is unlikely that many high-quality variables related to both the outcome and to treatment assignment will be collected in both a treatment and a comparison group in an observational study (or in a post-hoc analysis like this). Rather, it is more likely that simple demographic variables such as age, race, and gender might be collected, more to allow for subgroup comparison than to actually assess the equivalence of groups.

Beyond the simple mechanical problems associated with matching individuals with a limited set of covariates are more theoretical problems. It is exceedingly difficult to believe that the groups are equivalent on all of the unmeasured variables of interest when only a handful of

demographic variables are used in the matching process. In this example, there was a host of possible factors that could have been the “causal” variable that was associated with test performance. Variables such as a kindergarten pretest, socioeconomic status indicators, and a bevy of other predictors could have allowed for a better separation of the treatment and control groups through the propensity scoring; in addition, these variables would have provided the researcher with greater confidence that the “key” variables of interest for equalizing groups were being measured. With the limited variables in this example, it is very likely that the nonequivalent groups were significantly different according to an unmeasured variable that was related to the achievement outcome of interest.

One interesting finding pertains to the quality of the matching that occurred in these data. When conservative matching criteria were used, the results were generally slightly better, and less prone to the egregious errors outlined above. It must be noted, however, that by using only the subset of the treatment group in the program impact estimation that was most similar to the comparison group, external validity was limited. Clearly, it is imperative that the impact estimates be internally valid as a precursor to any external validity issue, but the issue of external validity may be a key one for policy decision-making.

It is attractive to consider the hypothetical increased power that would be associated with an increased sample size according to less restrictive propensity matching criteria. Only in Grade 3 (in matching without replacement) did it appear that additional power was noted. Findings became increasingly significant as the generosity of the matching criteria increased. However, as indicated earlier, based on the fact that the hypothetical increased power also wrongly identified program impact estimates that were *negative* and statistically significant, the increased power associated with generous matching criteria may have caused more harm than good.

In spite of the drawbacks associated with estimating propensity scores with limited covariates, program impacts from the created observational study were estimated. The gold standard of the randomized experiment was held as the “true” outcome, and in general, the propensity score program impact estimates did not accurately uncover the “true” program impact. The propensity score program estimates most commonly committed type II errors and did not reject the null hypothesis of no difference when in fact there was one. And sometimes, when the propensity score impact estimates were indeed significant, they were found to be *negative* when the true program impact was actually positive! This happened more frequently when matching occurred between treatment and control conditions with replacement. As this appears to be the more egregious error, it appears that matching without replacement might be a better solution in cases where the data are as sparse as those represented in this paper.

A serious limitation in this study can be found in the analysis under matching with replacement. Since certain students served as matches for multiple pairs, this analytic framework violated the nonindependence assumption. Agodini and Dynarski (2001) used a bootstrapping procedure to adjust the standard errors of their estimates, and in future work a similar analysis will be performed on these data to produce more confidence in these results. In addition, this analysis did not attempt to capture the nonindependence of students nested within the same school in the nonequivalent impact analysis. The assumption in these analyses was that the random effect for each pair captured all of the extraneous noise in the impact estimates. However, future researchers might want to empirically test this assumption by including additional random effects for schools in the nonequivalent group impact analyses.

In addition, the analysis performed on these data assumed that the effect of the small class size program was not additive across years, and that the effect was uncorrelated across the

four subject areas. Future analyses comparing these experimental and nonexperimental estimates might want to explicitly test whether different effect specifications are more or less sensitive to how the data are analyzed (i.e., from the experimental versus the observational framework).

In the end, the difficulties that arise when trying to create casual impact estimates from nonexperimental data are considerable, and may be impossible. In cases where

1) limited demographic data are available for the treatment and control groups, and

2) there are markedly fewer individuals in the control condition than the treatment condition,

it may be best for researchers to admit that with the lemons of observational data, they cannot create “causal” lemonade.

## References

- Agodini, R., & Dynarski, M. (2001). *Are experiments the only option? A look at dropout prevention programs*. Princeton, NJ: Mathematica Policy Research, Inc.
- Boruch, R. F. (1977). *Randomized experiments for planning and evaluation. A practical guide*.
- Boyd-Zaharias, J., Achilles, C. M., & Cain, V. A. (1995). The effect of random class assignment on elementary students' reading and mathematics achievement. *Research in the Schools*, 2(2), 7-14.
- Dehejia, R. H., & Wahba, S. (1998). *Propensity Score Matching for Non-Experimental Causal Studies*. National Bureau of Economic Research Working Paper #6829.
- Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94(448), 1053-1062.
- Fraker, T., & Maynard, R. (1987). Evaluating comparison group designs with employment-related programs. *Journal of Human Resources*, 22, 194-227.
- Friedlander, D., & Robins, P. (1995). Evaluating program evaluations: New evidence on commonly used nonexperimental methods. *American Economic Review*. 85(4), 923-937.
- Glass, G. V., Cahen, L. S., Smith, M. L., & Filby, N. N. (1982). *School class size: Research and policy*. Beverly Hills, CA: Sage.
- Glass, G. V., & Smith, M. E. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, 1, 2-16.
- Hanushek, E. A. (1999). The evidence on class size. In S. E. Mayer & P. E. Peterson (Eds.), *Earning and learning: How schools matter* (pp. 131-168). Washington, DC: Brookings Institution Press.

- Hedges, L. V., & Stock, W. (1983). The effects of class size: An examination of rival hypotheses. *American Educational Research Journal*, 20, 63-85.
- Krueger, A. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, 114(2), 497-532.
- LaLonde, R. (1986). Evaluating the econometric evaluations of training programs. *American Economic Review*, 76, 604-620.
- Mosteller, F., Light, R. J., & Sachs, J. A. (1996). Sustained inquiry in education: Lessons learned from skill grouping and class size. *Harvard Educational Review*, 66, 797-842.
- Nye, B. A., Hedges, L. V., & Konstantopolous, S. (1999). The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. *Educational Evaluation and Policy Analysis*, 21(2), 127-142.
- Parsons L. S. (2001). *Reducing bias in a propensity score matched-pair sample using greedy matching techniques*. Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc. Retrieved April 4, 2008, from <http://www2.sas.com/proceedings/sugi26/p214-26.pdf>
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.
- Victor, T. W. (2007). *A Monte Carlo evaluation of different ways to estimate treatment effect sizes in quasi-experiments and observational studies*. Dissertation defense presentation at the University of Pennsylvania, Graduate School of Education. April 23, 2007.
- Wilde, E. T., & Hollister, R. (2002). *How close is close enough? Testing nonexperimental estimates of impact against experimental estimates of impact with education test scores as outcomes*. (Discussion Paper No. 1242-02). Institute for Research on Poverty.

**APPENDIX**

Table 1:  
Grade 1 Descriptive Statistics

	STAR Control (Regular Class)					STAR Treatment (Small Class)					Non Equivalent Comparison (Regular Class)					
	N	Min	Max	Mean	Std	N	Min	Max	Mean	Std	N	Min	Max	Mean	Std	
Grade 1	Female		0	1	0.49	0.50		0	1	0.49	0.50		0	1	0.53	0.50
	Black		0	1	0.37	0.48		0	1	0.33	0.47		0	1	0.06	0.23
	Other		0	1	0.01	0.09		0	1	0.01	0.08		0	1	0.00	0.06
	SAT Reading Scale Score	2175	408	651	513.70	53.93	1607	417	651	529.98	56.22	831	410	651	534.27	49.96
	SAT Math Scale Score		408	676	525.98	41.57		425	676	538.90	44.14		425	676	547.02	40.34
	SAT Listening Scale Score		481	685	564.15	32.23		490	708	572.50	34.60		504	708	579.83	32.85
	SAT Study Skills Scale Score		317	601	507.40	53.71		381	601	523.48	52.28		410	616	541.89	46.19
Grade 2	Female		0	1	0.49	0.50		0	1	0.49	0.50		0	1	0.53	0.50
	Black		0	1	0.36	0.48		0	1	0.33	0.47		0	1	0.05	0.23
	Other		0	1	0.01	0.08		0	1	0.01	0.07		0	1	0.00	0.05
	SAT Reading Scale Score	2027	474	732	580.05	45.16	1774	470	732	590.69	46.05	837	477	732	595.82	45.40
	SAT Math Scale Score		441	721	578.11	43.35		455	721	586.78	45.74		475	721	594.19	43.49
	SAT Listening Scale Score		510	740	592.96	34.46		513	740	600.03	35.36		520	740	608.46	34.31
	SAT Study Skills Scale Score		451	672	579.89	49.90		465	672	590.57	50.77		468	707	596.71	52.80
Grade 3	Female		0	1	0.49	0.50		0	1	0.50	0.50		0	1	0.53	0.50
	Black		0	1	0.34	0.48		0	1	0.31	0.46		0	1	0.05	0.22
	Other		0	1	0.01	0.07		0	1	0.01	0.08		0	1	0.00	0.07
	SAT Reading Scale Score	1778	508	775	612.73	37.55	1896	511	775	621.28	39.23	653	494	753	619.67	38.00
	SAT Math Scale Score		519	774	616.04	38.99		487	774	622.87	40.20		477	752	623.28	38.13
	SAT Listening Scale Score		537	756	624.16	31.26		524	779	626.73	33.03		551	779	632.10	30.99
	SAT Study Skills Scale Score		494	740	608.35	43.63		499	740	618.05	45.20		477	740	614.91	43.47

Note: Indicated *n* applies for all variables within a treatment group within a given grade

Table 2  
Experimental Impact Estimates of Small Class Size by Grade

	Effect	SAT Reading Scale Score	SAT Math Scale Score	SAT Listening Scale Score	SAT Study Skills Scale Score
Grade 1	Intercept	598.746*** (14.426)	555.292*** (12.45)	583.776*** (9.639)	592.031*** (14.711)
	Female	8.71*** (1.533)	-1.199 (1.14)	-0.137 (0.926)	6.346*** (1.522)
	Age (in days)	-0.03*** (0.005)	-0.005 (0.003)	-0.002 (0.003)	-0.029*** (0.005)
	Black	-17.324*** (2.894)	-23.341*** (2.154)	-19.041*** (1.748)	-14.766*** (2.873)
	Other	-3.371 (9.29)	-9.815 (6.919)	-16.684* (5.616)	0.477 (9.224)
	Small Class	14.174*** (2.737)	12.474*** (2.493)	7.321** (1.901)	14.172*** (2.823)
	Grade 2	Intercept	676.23*** (10.402)	623.562*** (10.81)	641.323*** (8.134)
Female		7.071*** (1.27)	0.638 (1.236)	-3.229** (0.97)	3.846* (1.438)
Age (in days)		-0.049*** (0.003)	-0.027*** (0.003)	-0.018*** (0.003)	-0.057*** (0.004)
Black		-14.241*** (2.558)	-22.227*** (2.492)	-20.398*** (1.955)	-12.579*** (2.897)
Other		16.36 (8.613)	13.455 (8.389)	-1.77 (6.584)	1.06 (9.757)
Small Class		9.624*** (2.33)	9.307** (2.49)	6.055* (1.841)	9.79** (2.577)
Grade 3	Intercept	694.407*** (8.674)	674.91*** (9.039)	671.555*** (7.488)	691.899*** (10.109)
	Female	5.335*** (1.116)	-1.678 (1.125)	-4.421*** (0.926)	5.96*** (1.296)
	Age (in days)	-0.043*** (0.003)	-0.033*** (0.003)	-0.02*** (0.002)	-0.046*** (0.003)
	Black	-14.125*** (2.24)	-15.492*** (2.26)	-16.716*** (1.859)	-12.319*** (2.602)
	Other	6.71 (7.461)	6.516 (7.526)	-19.919* (6.193)	9.531 (8.666)
	Small Class	6.678** (2.006)	5.286* (2.122)	1.258 (1.763)	7.842** (2.341)

Note:  $n = 3782, 3801, \text{ and } 3674$  for grades 1,2,3 respectively

Standard Errors in Parentheses

\*\*\* =  $p < .0001$

\*\* =  $p < .001$

\* =  $p < .05$

Table 3  
Impact Estimates from different Propensity Score Matching Criteria (with replacement)

	N <sup>a</sup>	SAT Reading Scale Score	SAT Math Scale Score	SAT Listening Scale Score	SAT Study Skills Scale Score	
Grade 1	Experimental Impact	3782	14.17***	12.47***	7.32**	14.17***
	Propensity Match < .00001	1176	6.85*	3.15	1.16	-10.13***
	Propensity Match < .0001	1294	5.94*	2.42	1.20	-11.01***
	Propensity Match < .001	2202	6.21*	1.97	1.36	-10.33***
	Propensity Match < .01	3110	2.04	-3.02*	-0.65	-12.73***
	Propensity Match < .1	3214	2.88	-3.33*	-0.64	-11.82***
Grade 2	Experimental Impact	3801	9.62***	9.31**	6.06*	9.79**
	Propensity Match < .00001	1320	1.85	2.38	-0.85	-1.10
	Propensity Match < .0001	1478	1.21	2.02	-0.83	-0.82
	Propensity Match < .001	2556	0.63	-0.38	-1.30	-0.37
	Propensity Match < .01	3448	-0.04	-2.33	-1.10	-0.53
	Propensity Match < .1	3548	1.62	-1.76	-1.17	0.93
Grade 3	Experimental Impact	3674	6.68**	5.29*	1.26	7.84**
	Propensity Match < .00001	1250	9.77***	11.29***	2.39	11.41***
	Propensity Match < .0001	1498	9.88***	8.95***	2.63	11.04***
	Propensity Match < .001	2732	5.61***	4.10*	0.62	6.42**
	Propensity Match < .01	3644	6.98***	2.99*	0.05	8.25***
	Propensity Match < .1	3790	8.03***	4.71**	1.96	9.61***

Note: <sup>a</sup> N indicates the number of observations used in each estimation. For the propensity score estimates, N/2 indicates the number of pairs

\*\*\* = p < .0001

\*\* = p < .001

\* = p < .05

Table 4  
Impact Estimates from different Propensity Score Matching Criteria (without replacement)

	N <sup>a</sup>	SAT Reading Scale Score	SAT Math Scale Score	SAT Listening Scale Score	SAT Study Skills Scale Score	
Grade 1	Experimental Impact	3782	14.17***	12.47***	7.32**	14.17***
	Propensity Match < .00001	840	5.79	2.93	1.62	-10.44**
	Propensity Match < .0001	890	5.43	2.47	1.43	-10.64**
	Propensity Match < .001	1278	4.17	1.73	-0.51	12.85***
	Propensity Match < .01	1512	4.20	0.69	-1.44	12.30***
	Propensity Match < .1	1598	4.40	0.17	-1.98	12.11***
Grade 2	Experimental Impact	3801	9.62***	9.31**	6.06*	9.79**
	Propensity Match < .00001	904	2.82	3.88	2.08	0.59
	Propensity Match < .0001	956	1.74	2.70	1.06	-0.04
	Propensity Match < .001	1328	3.48	2.32	0.44	1.95
	Propensity Match < .01	1598	3.67	0.37	-0.73	1.14
	Propensity Match < .1	1658	4.85*	0.95	-0.45	2.56
Grade 3	Experimental Impact	3674	6.68**	5.29*	1.26	7.84**
	Propensity Match < .00001	762	8.51*	8.50*	3.11	9.51*
	Propensity Match < .0001	840	9.13**	9.49**	3.86	10.07*
	Propensity Match < .001	1218	8.56**	6.40*	1.41	9.40**
	Propensity Match < .01	1300	9.07***	7.27**	1.91	10.12***
	Propensity Match < .1	1304	8.98***	7.40**	1.99	10.12***

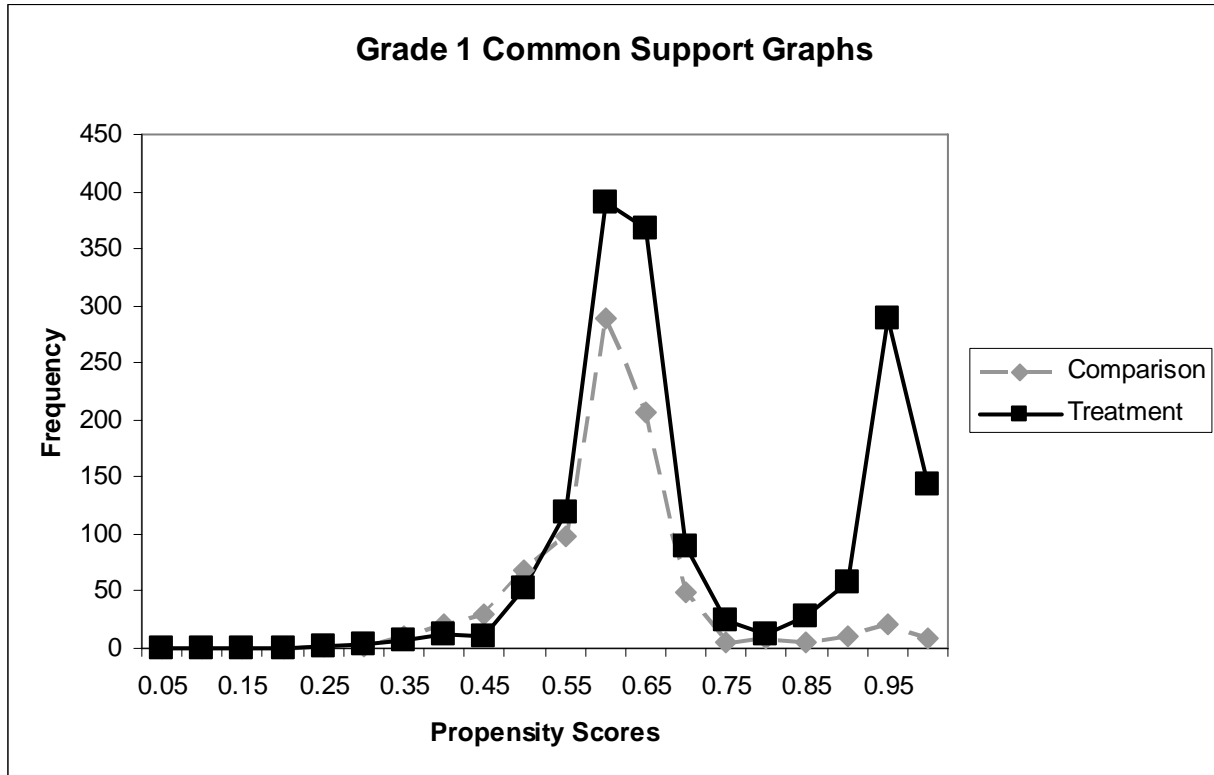
Note: <sup>a</sup> N indicates the number of observations used in each estimation. For the propensity score estimates, N/2 indicates the number of pairs

\*\*\* = p < .0001

\*\* = p < .001

\* = p < .05

**Figure 2.**



**Figure 3.**

