Contract No.: 2000-1337
MPR Reference No.: 8723-300

MATHEMATICA
Policy Research, Inc.

# Are Experiments the Only Option? A Look at Dropout Prevention Programs

***August 2001***

***Roberto Agodini***
***Mark Dynarski***

Submitted to:

Smith Richardson Foundation, Inc.
60 Jesup Road
Westport, CT 06880

Project Officer:
Phoebe H. Cottingham

Submitted by:

Mathematica Policy Research, Inc.
P.O. Box 2393
Princeton, NJ 08543-2393
(609) 799-3535

Project Director:
Mark Dynarski

## ABSTRACT

This article explores whether propensity score methods produce unbiased estimates of program impacts, by comparing experimental and propensity score impacts of dropout prevention programs. We find no consistent evidence that propensity score methods replicate experimental impacts in our setting. This finding holds even when the data available for matching are extensive. Our findings suggest that evaluators who plan to use nonexperimental methods, such as propensity score matching, need to carefully consider how programs recruit individuals and why individuals enter programs, as unobserved factors may exert powerful influences on outcomes that are not easily captured using nonexperimental methods.

# I. INTRODUCTION[1]

The validity of program impact estimates based on an experimental design is a powerful reason for the growing use of experimental designs. By creating a treatment group and a control group that are similar along all characteristics that affect outcomes, both observed and unobserved, experimental designs lead to a simple estimator of a program impact, which is the difference in average outcomes of the treatment and control groups. The standard error of the estimator also is easily estimated using the standard analytic formula.

Despite the appeal of experimental designs, they can be difficult to implement in many settings. For example, program operators often are reluctant to implement experimental designs when their programs are operating below capacity. Experimental designs are also difficult to implement when all individuals eligible for program services are affected by the treatment, which is true, for example, for statewide welfare reform efforts. Therefore, a nonexperimental design that produced valid impact estimates would have great appeal.

The findings of a recent article by Dehejia and Wahba (1999) suggest that a nonexperimental method that was developed some time ago, but was not often used to evaluate social programs—the propensity score method—may have potential to produce impacts similar to those that experiments would produce. In particular, Dehejia and Wahba showed that the propensity score method yielded impacts of a job-training program on earnings that were close to

[1]We thank Tim Novak for unparalleled research support. We also thank Rajeev Dehejia, Alan Krueger, David Myers, Don Rubin, Peter Schochet, and seminar participants at Mathematica for valuable comments. Last, but far from least, we thank the Smith Richardson Foundation for their generous support of this research; however, the findings and conclusions do not necessarily represent the official position or policies of the Smith Richardson Foundation. An earlier draft of this paper was presented at the 2001 annual meeting of the Econometric Society in New Orleans, LA. Correspondence can be directed to Roberto Agodini: phone (609) 936-2712 or email RAgodini@mathematica-mpr.com.

what experimental methods yielded. The propensity score method estimates impacts by comparing outcomes of a treatment group with outcomes of a select group of individuals who, on average, are similar to the treatment group along a wide array of observed characteristics. The select group of individuals (hereafter, comparison group) is selected from a sample of potential comparison group members using propensity scores—the probability of treatment status given the observed characteristics of treatment and potential comparison group members (Rubin 1973).

These findings raise the issue of whether propensity score methods can replicate experimental impacts of programs in other settings and for other outcomes. In theory, propensity score methods produce unbiased estimates of treatment effects if all the characteristics related to treatment status that are also related to outcomes are observed (Rosenbaum and Rubin 1983). In many social programs, treatment group members include individuals who were both eligible *and* interested in receiving program services. The characteristics used to determine eligibility are often known, making it possible to select a comparison group that is similar to the treatment group along these characteristics. However, the extent to which individuals are interested in program services can rarely be gauged. In situations where interest in receiving program services affects key outcomes, propensity score methods could lead to biased impact estimates.

In this study, we explore the general applicability of propensity score methods by comparing experimental and propensity score impacts of 16 dropout prevention programs on 4 student outcomes, including dropping out, absenteeism, educational aspirations, and self-esteem. We also estimate the standard error of propensity score impacts using a bootstrap approach, which allows us to assess the power of propensity score-based designs and to determine how well the standard analytic formula estimates the complex variances that arise when matching methods are used.

The analysis is based on data collected for the federal evaluation of the School Dropout Demonstration Assistance Program (SDDAP) and the National Education Longitudinal Study (NELS). From 1991 to 1995, the SDDAP funded programs throughout the United States designed to reduce dropping out among middle- and high-school students who were at risk of dropping out. The SDDAP data are ideal for this study because they contain more than 150 data items for two random samples of students who were at risk of dropping out *and* who were interested in receiving program services: (1) treatment groups that were offered program services, and (2) control groups that were not. These samples enable us to compute experimental impacts of each of the 16 programs. The SDDAP data also contain the same information for a third sample of students who were at risk of dropping out, but who did not have access to program services. This sample enables us to compute propensity score impacts based on comparison groups that were matched to the treatment groups using data items from identical questionnaires. The NELS also is useful for this study because, although it does not contain all the survey items available in the SDDAP data, it is one of the largest, nationally representative samples of students that is publicly available. These data enable us to compute propensity score impacts based on data that researchers who do not have access to primary data are likely to use to evaluate the effect of education programs, such as those funded by the SDDAP.

Two main findings emerged from this study. First, we find no consistent evidence that propensity score methods replicate experimental impacts of the SDDAP programs. Among the 16 programs for which we estimated impacts, there are only scattered instances in which the experimental and propensity score impacts are similar. Moreover, no patterns are evident in the results to suggest the types of programs for which propensity score methods may be more likely to replicate experimental impacts. This is true for propensity score impacts based on extensive data and those based on less extensive data that researchers are likely to have access to. Second,

our findings suggest that propensity score methods can replicate standard errors of experimental impacts. Moreover, they can be computed using the standard analytic formula.

The article is organized in the following way. Section II describes some common approaches to measuring program impacts and what the literature has found regarding the validity of these approaches. Section III outlines the goals of this study and describes the data used for the analysis. Section IV describes how propensity score matching was used to select comparison groups. Section V describes the methods used to estimate impacts and standard errors. Sections VI and VII present our results and conclusions, respectively.

## II. EVALUATION DESIGN OPTIONS

In theory, the impact of a program on its participants is the difference between the outcomes participants experience after they participated in the program, and the outcomes they would have experienced had they not participated in the program. The outcomes participants would have experienced had they not participated in the program are often referred to as the "counterfactual." It is not possible to observe the counterfactual. Therefore, the major challenge facing evaluations is how to estimate the counterfactual.

Some common approaches to estimating the counterfactual include pre-post designs, comparison group designs, and experimental designs. Pre-post designs measure outcomes of participants before and after they participate in the program, essentially using pre-program outcomes as an estimate of the counterfactual.

Though generally straightforward to implement, pre-post designs often can generate misleading impact estimates. Individuals may be maturing rapidly, and their outcomes may be quite different even a short time later, regardless of whether they participate in a program. This is especially true of children and youth. For example, a pre-post evaluation of a program

designed to increase school attendance faces the problem of attendance commonly decreasing as students mature. Unless the program's impact on attendance is dramatic, attendance rates before students participated in the program could be greater than attendance rates afterward. A pre-post design would yield a finding that the program *reduced* attendance, instead of *increasing* it, as expected.

Comparison group designs improve on pre-post designs by using a group that is "similar" to participants, so that the influence of trends and other factors on outcomes can be reduced. In the example just given, attendance rates for a suitably chosen comparison group may also show a decline, which allows the evaluator to infer whether the attendance decline of participants was lower than it would have been without the program. A lower rate of decline for participants relative to the comparison group would suggest a positive impact of the program on attendance.

Comparison group designs resolve some, but not all of the issues involved in measuring the counterfactual. The key unresolved issue is whether members of the comparison group are similar along all relevant dimensions to participants. For example, if the attendance program serves all the students in a school with poor attendance, the evaluator would need to draw a comparison group either from among students in that school who had better attendance, or from students who also had poor attendance in another school that did not operate an attendance program. Both options introduce the possibility of dissimilarity. Using the first option means that students with poor attendance will be compared to students with better attendance, who could well have different attendance trends. Using the second option means that students with poor attendance in one school will be compared to students with poor attendance in another school, whose treads may differ from the school where the program is being studied. The possibility of dissimilarity means that measured differences in outcomes between participants

and the comparison group will consist both of program impacts and the influence of other factors. Separating program impacts from other factors may be difficult in these cases.

Experimental designs resolve the dissimilarity problem by randomly assigning individuals who are eligible to participate in a program to one of two groups. The first group—often referred to as the "treatment" group—is allowed to participate in the program, whereas the second group—often referred to as the "control" group—is not. The impact of the program is estimated as the difference in outcomes between the treatment and control groups.

Experiments are appealing because they yield the ideal comparison group—a group that is similar to the program group in terms of dimensions that can be measured (such as age, sex, and race), as well as those dimensions (such as motivation and attitudes) that cannot be measured or that can be measured only at prohibitive cost. In the example of the attendance program, the two groups would be similar in terms of their motivation to do well in school—a characteristic that is difficult to observe directly, but which is likely to affect attendance.

The theoretical superiority of experimental designs over pre-post and comparison group designs does not mean that the latter necessarily gives false or misleading answers in practice. If unobserved factors do not influence outcomes much, or if similar comparison groups can be readily identified, comparison group designs may well yield results that are close to results from experimental designs. Furthermore, mounting an experiment presents its own challenges, including addressing the perceived ethical problem of not offering services to control group members, the need to ensure that programs have adequate numbers of applicants to create control groups without leaving program slots vacant, and the cost of monitoring to ensure that experimental integrity is maintained. If comparison group designs produce results (both impact estimates and their standard errors) that come "close enough" to those based on experimental

designs, the costs of mounting experiments could well outweigh the benefits of experimental results.

Beginning in the 1980s, researchers have examined whether comparison group designs can replicate results from experimental designs for a particular program. These researchers compared experimental impact estimates to impact estimates based on comparison groups. Experimental impacts were estimated as the difference in average outcomes between the treatment and control group. Comparison group impacts were estimated in a similar way, except that comparison groups were used instead of control groups. The comparison groups included individuals from other samples, such as the Current Population Survey, that were collected around the same time that treatment group members were offered services, and that contained information that also was available for treatment group members. To adjust for differences between the treatment and comparison groups that may have been related to outcomes, comparison group impacts were also estimated using subsets of individuals in these samples, such as those that were identical to the treatment groups along some of the program's eligibility criteria. To further adjust for important differences between the treatment and comparison groups, econometric techniques were often used to refine comparison group impacts.

The general findings were that comparison group designs do not come close to experimental designs and that they often yield highly misleading findings. Lalonde (1986) compared experimental impact estimates with those based on several comparison groups and found that the experimental and comparison group impact estimates were strikingly different. In particular, he found that the experimental impact estimate of the National Supported Work demonstration (NSW) on earnings was about $1,800, whereas comparison group estimates ranged widely from –$15,000 to $1,000. Fraker and Maynard (1987) and Friedlander and Robins (1995), who used a different data source and different methods for creating comparison groups, also found that

comparison group methods yielded inaccurate estimates of the effects of welfare reform programs on employment rates.

Using the same data as Lalonde, a recent article by Dehejia and Wahba (1999) showed that a nonexperimental method not often used by evaluators—propensity score matching—yields impact estimates that are close to those produced by an experimental design. Propensity score methods estimate impacts by comparing outcomes of program participants with outcomes of a select group of individuals who, on average, are similar to participants along all the characteristics that are related to the outcomes of interest (Rubin 1973).

Whereas Lalonde found that comparison group estimates of the NSW's impact on earnings varied widely, Dehejia and Wahba found that impact estimates of the program based on propensity score methods came fairly close to experimental impact estimates. In particular, for the subset of NSW treatment group members studied by Dehejia and Wahba,[2] the experimental estimate of the program's effect on earnings equaled $1,794, whereas impact estimates based on various approaches of propensity score matching ranged from $1,200 to $2,200. For one particular approach of propensity score matching—nearest neighbor matching, which we

[2]Dehejia and Wahba's sample includes two types of individuals from the NSW sample used by LaLonde. The first type includes individuals who were randomized midway through the program's intake period. The authors included these individuals because they had two years of pre-intervention earnings information. (Those who were randomized earlier only had one year of pre-intervention earnings information.) Dehejia and Wahba wanted to include individuals who had two years of pre-intervention earnings because Ashenfelter (1978) and Ashenfelter and Card (1985) showed that using more than one year of pre-intervention earnings is critical for accurately estimating the effect of training programs. The second type includes individuals who were randomized later in the program's intake period. The authors included these individuals because they also had two years of pre-intervention earnings. However, among these individuals, the authors only included those who were unemployed prior to randomization. The authors do not explain why the latter restriction was imposed. Whatever the case, the pre-intervention characteristics of the treatment and control group members in their sample are similar.

describe later in the article and use to select our comparison groups—the propensity score impact, at $1,691, was very close to the experimental impact of $1,794.

Dehejia and Wahba's findings raise important issues about the general applicability of propensity score methods, such as whether it can replicate experimental impacts of programs in other settings and for other outcomes. The theoretical properties of the propensity score method ensure that, if unobservable characteristics do not influence outcomes, matching individuals based on their propensity score is equivalent to using an experimental design (Rosenbaum 1995). If unobservable characteristics influence outcomes, propensity score methods may yield different impact estimates than those experimental methods would produce. However, the difference depends on the extent to which program participants and the comparison group differ along unobservable characteristics. Dehejia and Wahba's results suggest that, at least for the subset of NSW treatment group members they studied, the extent to which unobservable characteristics mattered was small enough for propensity score methods to do well in replicating experimental findings.[3]

[3]Smith and Todd (2000) concluded that unobservable characteristics did not matter much in Dehejia and Wahba's subset of the NSW data because Dehejia and Wahba's subset only included individuals who were randomized later in the program's intake period, if they were unemployed prior to randomization (Footnote 1 describes who was included in Dehejia and Wahba's sample). In fact, Smith and Todd showed that propensity score methods do not replicate experimental impacts when the individuals that Dehejia and Wahba excluded are included in the analysis.

## III. THE GOALS OF THIS STUDY AND THE DATA

This study explores the applicability of propensity score methods by addressing the following questions:

- How well do propensity score methods replicate experimental impacts of programs in other settings and for other outcomes?
- How well do propensity score methods replicate experimental impacts when based on limited information commonly available from public-use data sets, rather than from more extensive data sets?
- How precise (standard errors) are impact estimates based on propensity score methods—an issue not considered in previous applications of propensity score matching?

The analysis is based on data collected for the federal evaluation of the School Dropout Demonstration Assistance Program (SDDAP) and data collected for students who participated in the National Education Longitudinal Study (NELS).

### A. The School Dropout Demonstration Assistance Program

Operating between 1991 and 1996, the SDDAP funded two types of programs throughout the United States that were designed to reduce dropping out among both middle- and high-school students: (1) targeted, and (2) restructuring. The targeted programs offered services to *students* who met eligibility criteria that programs set for themselves, which included students being overage for grade, having low grades or test scores, having frequent absences, having a history of disciplinary incidents, or having alcohol or substance-use problems. However, the criteria were general guidelines, and most programs retained substantial discretion to serve any student that they deemed in need of, or able to benefit from, services. For example, some programs served many students who were already dropouts. Students often were referred to programs by teachers, counselors, or other school staff, or they expressed interest on their own. Analyses of

student characteristics at baseline indicated that nearly all students served by the programs were either dropouts, or at risk of dropping out, using conventional risk factors (Gleason and Dynarski 1994).

The restructuring programs promoted *schoolwide* reform designed to reduce dropping out. Unlike the targeted programs, the restructuring programs tried to put in place services and structures designed to affect all students in a school and, ultimately, lower the school's dropout rate. The services and structures were intended to be comprehensive and included new curriculum approaches, changes in school governance, expanded teacher training and development, and expanded health and social services. The schools that were part of restructuring programs generally were those with significant numbers of at-risk students, which was one of the criteria for receiving a grant and which was verified by analyses of student characteristics (Gleason and Dynarski 1995). However, students did not have to be dropouts or at risk of dropping out to be in the study sample (as were nearly all students served by targeted programs). Generally, students were sampled randomly from lists of seventh graders in restructuring middle-schools, or from lists of tenth graders in restructuring high-schools. Ninth graders were sampled in one of the four school districts because the restructuring effort there focused on the ninth grade.

A total of 85 programs, both targeted and restructuring, were funded by the SDDAP. Of the 85 programs, 65 were established in September 1991 and the remainder in 1992. Mathematica Policy Research, Inc., selected 20 of the 65 programs funded in September 1991 to be part of the federal evaluation of the SDDAP. Of these, 16 were targeted programs and 4 were restructuring programs. Selected programs generally were those offering innovative services or activities that were able to meet the evaluation's research requirements, including being able to conduct random assignment and serving enough students to generate adequate sample sizes.

The targeted programs were evaluated using an experimental design, whereas the restructuring programs were evaluated using a comparison group design. For each targeted program, the experimental design essentially involved comparing outcomes of two random samples of students who met the program's eligibility criteria and who were interested in receiving programs services, where one (the treatment group) was offered services, and the other (the control group) was not. More details about the comparison group design used to evaluate the restructuring programs are provided in the next section.

## B. Treatment Groups

The treatment groups used in this study include students who were randomly assigned to each of the 16 targeted programs, of which 8 were middle-school programs and the other 8 were high-school programs. As was mentioned above, treatment group members were either dropouts, or at risk of dropping out. The size of the treatment groups ranged from 77 to 393, with a total of about 3,000 treatment group members. About half of this sample was selected in the fall of 1992, and the other half in the fall of 1993.

Baseline and follow-up data were collected for treatment group members. The follow-up data were collected at one-year intervals. Three followups were conducted for the 1992 cohort, whereas two followups were conducted for the 1993 cohort. As such, four data points (baseline plus three followups) are available for the 1992 cohort, whereas three data points (baseline plus two followups) are available for the 1993 cohort. We analyze outcomes of treatment group members two years after baseline because this information is available for both cohorts.

The baseline and follow-up data come from school records and questionnaires. The school records data included 8 items and the questionnaire data included more than 150 items. Together, the school records and questionnaire data provide extensive information about each

student's characteristics and outcomes at baseline, as well as how those characteristics and outcomes changed during each subsequent followup.

For treatment groups of some programs, however, some data items were not collected because they were not available. For example, reading and math test scores were not available for some treatment groups. Therefore, as we explain later, comparison groups for some treatment groups were selected using fewer characteristics than was the case for other treatment groups.

## C. Potential Comparison Group Members

Two comparison groups were selected for each of the 16 treatment groups. This section describes the two samples of students from which comparison groups were selected. It also discusses the advantages and disadvantages of selecting comparison groups from each sample.

***SDDAP Comparison School Students.*** The first set of comparison groups was selected from students who attended the comparison schools used to evaluate the SDDAP restructuring programs. As was mentioned above, a comparison group design was used to evaluate the restructuring programs. Essentially, this involved comparing outcomes of two random samples of students, where one sample attended a school that operated a restructuring program and the other attended similar schools that were in the same district but which did not operate a restructuring program (hereafter, comparison school). As was mentioned above, four restructuring programs were included in the evaluation, with one comparison school selected for each restructuring school.[4] We selected the first set of comparison groups from the pooled sample of SDDAP comparison school students. The pooled sample contains about 3,000

[4]In one district, three middle schools were selected as comparison schools for two middle schools that operated a restructuring program.

students. Outcomes of these students were measured two years after baseline, which is consistent with the point at which outcomes were measured for treatment group members.

There are two advantages of selecting comparison groups from this sample of students. First, by design, these students were at risk of dropping out, and they attended schools where many students were at risk of dropping out. Second, the same school records and questionnaire data collected for the treatment groups were also collected for students who attended the SDDAP comparison schools. Therefore, this sample contains many students who met the eligibility criteria met by the treatment groups, and can be further matched to the treatment groups using the same extensive information that was collected for the treatment groups.

A potential disadvantage of selecting comparison groups from SDDAP students is that they did not attend schools in the same district as the schools attended by treatment group members. SDDAP students attended school in Grand Rapids, Michigan; Dallas, Texas; Phoenix, Arizona; or Santa Ana, California. None of the treatment group students attended school in these cities. If school-location characteristics were important determinants of the outcomes analyzed, the school-location differences between treatment and SDDAP students may affect the propensity score method's ability to replicate experimental impacts in our setting.

***NELS Students.*** Our second set of comparison groups was selected from students who participated in the NELS. The base-year NELS survey was conducted in 1988 and contained a nationally representative sample of eighth graders.[5] Follow-up surveys were conducted in 1990, 1992, 1994, and 2000. While respondents were of school age—which includes the 1988, 1990, and 1992 surveys—information was collected from students, one of their parents, two of their teachers, and their school's administrator. Some students, although of school age, were not in

[5]See Spencer et al. (1990) for more information about the base-year NELS sample design.

school during the 1990 and 1992 surveys because they dropped out. To understand why these students dropped out, information related to dropping out was collected from them. After respondents should have graduated from high school—which includes the 1994 and 2000 surveys—information was collected only from respondents and not from others who were previously surveyed, such as parents. High school transcripts of respondents were also collected.

Our NELS sample includes: (1) students who participated in the 1988 survey, and (2) students who participated in the 1990 survey. Taken together, this sample contains about 28,000 students.[6] However, it is not a sample of unique students; instead, it contains many students twice because there are many students who participated in both the 1988 base-year survey and the 1990 follow-up survey, which are the criteria we used to define our NELS sample. We included these students twice because whether one of these students is a suitable comparison group member may depend on the point in time at which we measure their characteristics.[7] As was the case for SDDAP students, outcomes of NELS students were measured two years after baseline, which is consistent with the point at which outcomes were measured for treatment group members.

[6]The 1988 survey included about 25,000 students and the 1990 survey included about 19,000 students, suggesting that our NELS sample should include about 44,000 students because it included both 1988 and 1990 sample members. The reason it includes 28,000 students, or 16,000 fewer students, is that 8,000 students did not have outcome data, and the remaining 8,000 did not have information that indicated the urbanicity of their school—one of the characteristics we want to use in the matching process.

[7]For example, suppose that we are looking for a comparison group member who should be in the tenth grade, but instead is in the ninth grade because he or she was left back or is behind grade-level. Also, suppose that a NELS student was in the eighth grade and on grade-level during the base-year survey, but in the ninth grade and behind grade-level during the first follow-up survey. This NELS student would not be a suitable comparison group member if we examined his or her grade-level during the base-year survey, but would be a suitable comparison group member if we examined his or her grade-level during the first follow-up survey.

There are two advantages of selecting comparison groups from this sample of students. First, it is a public-use data set of students who attended school at roughly the same time as the treatment groups. Therefore, these are the type of data that researchers who lack access to primary data are likely to use when evaluating education programs such as those funded by the SDDAP. Second, although the data set does not contain enough students who attended schools in the same districts as the treatment groups, it does contain enough students who attended schools in areas that are similar according to level of urbanicity.[8] Therefore, these data allow us to use school-location characteristics in the matching process. Including school-location characteristics in the matching process may be important, as Heckman et al. (1998) found in their study of job-training programs that comparison groups selected from the same areas as the treatment groups were more likely to experience the same outcomes as control groups from the same areas.

A potential disadvantage of selecting comparison groups from the NELS is that we cannot use in the matching process all the characteristics that are available for the treatment groups. By design, the questionnaire administered to the treatment groups used many of the same items as the NELS questionnaire. However, the questionnaire administered to the treatment groups contained some characteristics that could be used to select comparison groups but which the NELS questionnaire did not include.

[8]NELS students were classified as attending school in one of seven types of areas: (1) large central city, (2) mid-size central city, (3) small town, (4) urban fringe of large city, (5) urban fringe of mid-size city, (6) large town, or (7) rural.

# IV. HOW COMPARISON GROUPS WERE SELECTED AND WHAT THEY LOOK LIKE

A straightforward way to select a comparison group is, for each treatment group member, to select a potential comparison group member who is identical along each characteristic that affects outcomes. This approach would ensure that the selected comparison group experiences the outcomes the treatment group would have had they not been exposed to program services.

The problem with this approach is that it may be difficult to find a comparison group member for each treatment group member when many characteristics are used in the matching process. For example, if 10 dichotomous variables are used in the matching process, there are 1,024 possible values for the collection of variables.

## A. Propensity Score Matching

Rosenbaum and Rubin (1983) showed that propensity scores could be used to reduce the dimensionality problem. The propensity score is a scalar that equals the probability of treatment status given the observed characteristics of treatment and potential comparison group members. In particular, they showed that, in situations where the outcome is independent of treatment status given observed characteristics, then the outcome is also independent of treatment status given the propensity score. Matching individuals using propensity scores produces a comparison group that is similar, on average, to the treatment group along observed characteristics.

We used propensity scores to select comparison groups for each of the treatment groups, according to the following three steps. First, a logit model with dependent variable that indicates treatment status and independent variables that represent student characteristics was estimated using treatment and potential comparison group members. Second, parameter estimates of the logit model, along with each student's values for the respective independent variables, were used to assign to each treatment and potential comparison group member their likelihood of being a

treatment group member, or their propensity score. Third, for each treatment group member, the potential comparison group member with the closest absolute propensity score, or the nearest neighbor, was selected.

A point worth emphasizing about the third step is that the selection process was done with replacement, so that a potential comparison group member could have been matched to several treatment group members. Research has shown that impacts based on a comparison group selected with replacement can be similar to those experimental methods would produce, whereas impacts based on a comparison group selected without replacement may differ (Dehejia and Wahba 1999). Selecting with replacement is particularly important in situations where there are few similar, potential comparison group members.[9]

### B. Tests Used to Assess the Similarity of the Treatment and Comparison Groups

Rosenbaum and Rubin (1983) also showed that propensity score methods can be used to estimate treatment effects if two conditions are satisfied: (1) all the characteristics related to treatment status that are also related to outcomes are observed, and (2) treatment and comparison group members with similar propensity scores are balanced along these characteristics. The latter condition means that the logit model must produce an estimate of the propensity score such that, at each value of the estimated propensity score, the characteristics of treatment and comparison group members are similar.

[9]For each treatment group, we also estimated impacts based on a comparison group that includes all potential comparison group members whose propensity score falls within the minimum and maximum value of the treatment group's propensity score distribution. These results were similar to the results based on the nearest-neighbor comparison group. This is consistent with Smith and Todd (2000), who found that propensity score impacts are not sensitive to the way in which the propensity score is used to select comparison groups.

Generally speaking, we tested whether our comparison groups satisfy the second condition by comparing the characteristics of treatment and comparison group members with similar propensity scores, as the second condition indicates. More specifically, we first assigned treatment and comparison group members to strata, where each stratum included treatment and comparison group members whose average propensity score was not statistically different. The strata were defined by ranking treatment and comparison group members according to their propensity scores. Beginning with the observation with the highest propensity score and working backward, observations were dropped until the average propensity score of treatment and comparison group members among the observations that remained was, according to a t-test, not statistically different at the 0.05 level of confidence. The observations that remained were assigned to the first stratum. The average propensity score of treatment and comparison group members who were dropped in the previous step was then compared. If it was not statistically different, we considered the strata—which in this case equals 2—to be defined. If it differed, additional strata were defined, until the average propensity score of treatment and comparison group members in each stratum was not statistically different. Within each stratum, when then conducted an F-test of the similarity of the collection of matching characteristics across treatment and comparison group members.[10] If the F-test in each stratum failed to detect a difference at the 0.05 level of confidence, we concluded that our comparison group satisfied the second condition. If any one of the F-tests detected a difference, we respecified the logit model

[10]F-tests were estimated using a regression model with dependent variable that indicates treatment status and independent variables that represent the characteristics used in the matching process.

by adding higher-order or interaction terms and reselected a comparison group until all the F-tests failed to detect a difference.[11]

### C. Characteristics Used in the Matching Process

Our goal was to select comparison groups that are similar to their respective treatment groups along all the characteristics that affect the outcomes for which we compute impacts. This would ensure that the comparison groups experience the outcomes their respective treatment groups would have experienced had they not been exposed to program services. In other words, it would ensure that the outcomes of the comparison groups are a reliable estimate of the counterfactual.

Our approach for meeting this goal had two components. First, we included in the matching process the characteristics used to determine program eligibility. Second, we identified the characteristics that the literature indicates are related to dropping out, determined which of those characteristics were related to the dropout status of our treatment groups at baseline, and also used those characteristics in the matching process. The idea behind the second component is that we also wanted to use in the matching process those characteristics that are related to dropping

[11]We also used the test often used by the literature to determine whether our comparison groups satisfied the second condition (see, for example, Dehejia and Wahba 1998). In particular, we first ranked the collection of treatment and comparison group members in ascending order according to their propensity scores. Individuals were then broken up into propensity score strata with imposed cut-offs, instead of using the data to determine the cut-offs, as the test above does. To ensure that the statistical tests in each stratum had enough power, the number of observations in each stratum was equal to at least twice the number of variables included in the logit model. Within each stratum, we then conducted two statistical tests. The first was a t-test of the similarity of the average propensity score of treatment and comparison group members. The second was an F-test of the similarity of the characteristics of treatment and comparison group members. In most cases, this approach also suggested that our comparison groups satisfy the second condition. In other words, within each stratum, the average propensity score of the treatment and comparison groups was not statistically different, nor were their characteristics. The results of these statistical tests are available upon request.

out among the students for which we *actually* are selecting comparison groups—that is, for our treatment groups. Also, we want to use in the matching process those characteristics that are related to dropping out *before* the treatment groups were exposed to program services—that is, at baseline.

To determine which characteristics are related to dropping out among our treatment groups, we used a regression model to analyze the dropout status of our treatment groups at baseline. We also analyzed baseline values of the other three outcomes—absenteeism, educational aspirations, and self-esteem—for which we compute impacts. The models included characteristics that the literature indicates are related to dropping out and that are available for our treatment groups.[12] It also included several characteristics that the literature did not indicate are related to dropping out, but which we thought may be important to use in the matching process. For example, we included in the regression model the extent to which our treatment groups participated in extracurricular activities, because this characteristic may be correlated with the extent to which students feel a sense of belonging to their school, which, in turn, may discourage dropping out.

Table 1 reports all these characteristics, of which there are 32, and which of them are related to at least one of the baseline outcomes of our treatment groups.[13] The results are reported separately for the pooled middle-school treatment groups and the pooled high-school treatment

[12]Characteristics that the literature indicates are related to dropping out, but that either are not available for our treatment groups or had a high proportion of missing values, include curricular track, income, having to care for a child, employment status, measures of socioeconomic status other than mother's education, religiosity, and neighborhood socioeconomic status.

[13]Four of the characteristics are baseline values of the outcomes for which we compute impacts. When analyzing the baseline value of a particular outcome, we included in the analysis baseline values of the other outcomes.

TABLE 1

CHARACTERISTICS RELATED TO BASELINE OUTCOMES

| | Treatment Groups | | |
|---|---|---|---|
| | Middle School | High School | Either Group |
| **Baseline Outcome** | | | |
| Dropout | X | | X |
| Absenteeism | X | | X |
| Educational aspirations | X | | X |
| Self-esteem | X | | X |
| **Student Characteristics** | | | |
| Age[a] | X | X | X |
| Sex | X | X | X |
| Race/ethnicity | X | X | X |
| Reading test score | X | X | X |
| Math test score | X | | X |
| Time spent reading for fun[a] | X | X | X |
| Time spent watching TV[a] | X | X | X |
| Mother's education | X | X | X |
| Father's education | | | |
| Ever dropped out[a] | X | X | X |
| Time spent doing homework | X | X | X |
| Sibling ever dropped out | | X | X |
| Number of schools attended since 1st grade[a] | X | | X |
| Number of siblings[a] | | | |
| Talk about school with parents[a] | X | X | X |
| Ever skip school[a] | | X | X |
| Ever late for school[a] | X | X | X |
| Active in extracurricular activities[a] | X | | X |
| Does not live with both parents[a] | X | X | X |
| On public assistance | X | | X |
| Mostly speaks another language | | | |
| Overage for grade | | | |
| Low course grades | X | | X |
| Discipline problems | X | | X |
| Locus of control | X | X | X |
| Has own child | | | |
| **School and Neighborhood Characteristics** | | | |
| Level of Urbanicity | X | X | X |
| School Climate[a] | X | X | X |

SOURCE: Authors' calculations based on data collected for the federal evaluation of the School Dropout Demonstration Assistance Program.

[a]Characteristic the literature did not indicate is related to dropping out, but that we thought may be important to use in the matching process.

X = Significantly different from zero at the 0.05 level, two-tailed test

groups. The characteristics that are related to at least one of the baseline outcomes of either treatment group also are reported.

Among the 32 characteristics, the results indicate that 27 are related to at least one of the baseline outcomes of either treatment group. The related characteristics include baseline values of the outcomes for which we compute impacts, many student characteristics, and a few school and neighborhood characteristics.

We tried to use in the matching process all 27 related characteristics, plus another characteristic that was not related—overage for grade. We included "overage for grade" because it was one of the main criteria the targeted programs used to determine if a student was eligible for services.

For a number of reasons, the matching process sometimes could not use all the related characteristics. First, as we mentioned above, some of the characteristics were not collected for some of the treatment groups or for a potential comparison group. In particular, characteristics that were not collected for some of the treatment groups include absenteeism, reading test scores, and math test scores, mostly due to their unavailability in school records. Similarly, several characteristics were not collected for NELS students, including math and reading test scores, time spent reading for fun, whether a sibling had dropped out, active in extracurricular activities, discipline problems, and absenteeism.[14]

Another reason why the matching process sometimes could not use all the related characteristics is that, even though a characteristic was collected for both a treatment and a

[14]Actually, some of these characteristics are available for NELS students; however, they are not identical to those available for the treatment groups. For example, math and reading test scores are available for NELS students. However, the math and reading tests that were administered to NELS students are different than those that were administered to the treatment groups.

potential comparison group, there sometimes was no overlap between the two groups along that characteristic. In particular, school urbanicity is available for all the treatment groups and SDDAP comparison school students. However, half the treatment groups attended a school in a mid-size central city or a small town, whereas none of the SDDAP comparison school students attended a school in those areas. This was not an issue for NELS students because there were a significant number who attended school in each type of area where the treatment groups attended school.

Finally, the matching process sometimes could not use all the related characteristics because we could not select a comparison group that satisfied the second condition that Rosenbaum and Rubin (1983) showed must be satisfied when using propensity score methods to estimate treatment effects. In particular, we could not use certain characteristics because they resulted in a comparison group that, within each propensity score stratum, was not similar to its respective treatment group along those characteristics. Characteristics that often could not be used in the matching process included: mother's education, prior dropout status, baseline dropout status, and overage for grade. We tried to include these characteristics in different or more complex ways. For example, we tried to include in the logit model different types of variables for these characteristics, such as categorical variables with different cut-offs for the categories. We also tried to include higher-order terms for these characteristics and interact them with other characteristics. Despite our efforts, we could not use certain characteristics to select comparison groups for some of our treatment groups.

Tables A.1 and A.2 in the appendix report the characteristics used to select the SDDAP and NELS comparison groups, respectively, for each of the treatment groups. A maximum of 20 characteristics were available to select an SDDAP comparison group. However, only the SDDAP comparison group for the Miami Corporate Academy treatment group was selected

using all of the available characteristics; the rest were selected using between 13 and 19 characteristics. While 13 was the minimum number of characteristics that were used to select an SDDAP comparison group, it was the maximum number that was available to select a NELS comparison group. And, as was the case for the SDDAP comparison groups, only the NELS comparison group for the Miami Corporate Academy treatment group could be selected using all the available characteristics; the rest were selected using between 7 and 12 characteristics.

### D. Summary

Figure 1 summarizes the process we used to select comparison groups. It is worth emphasizing that this process make no use of outcome information. We know, within some degree of statistical precision, the "right" impacts—that is, the experimental impacts for each of the 16 programs. This makes it possible, at least in theory, to search for a comparison group that replicates the experimental impacts. However, such a search process would be of no help in designing an evaluation, which needs occur before outcomes are observed. Rubin (2001) emphasizes the importance of this point.

### E. Similarity of the Treatment and Comparison Groups

Table 2 reports statistics on the similarity of the matching characteristics across each treatment group and three groups. The first group is the randomly assigned control group used for the experimental evaluation of the dropout prevention programs. The second and third groups are the SDDAP and NELS comparison groups, respectively. The statistics in the table are estimated p-values from F-tests of the similarity of the collection of matching characteristics across the treatment and control/comparison groups. P-values were estimated using a regression model with dependent variable that indicates treatment status and independent variables that

FIGURE 1

PROCESS USED TO SELECT COMPARISON GROUPS

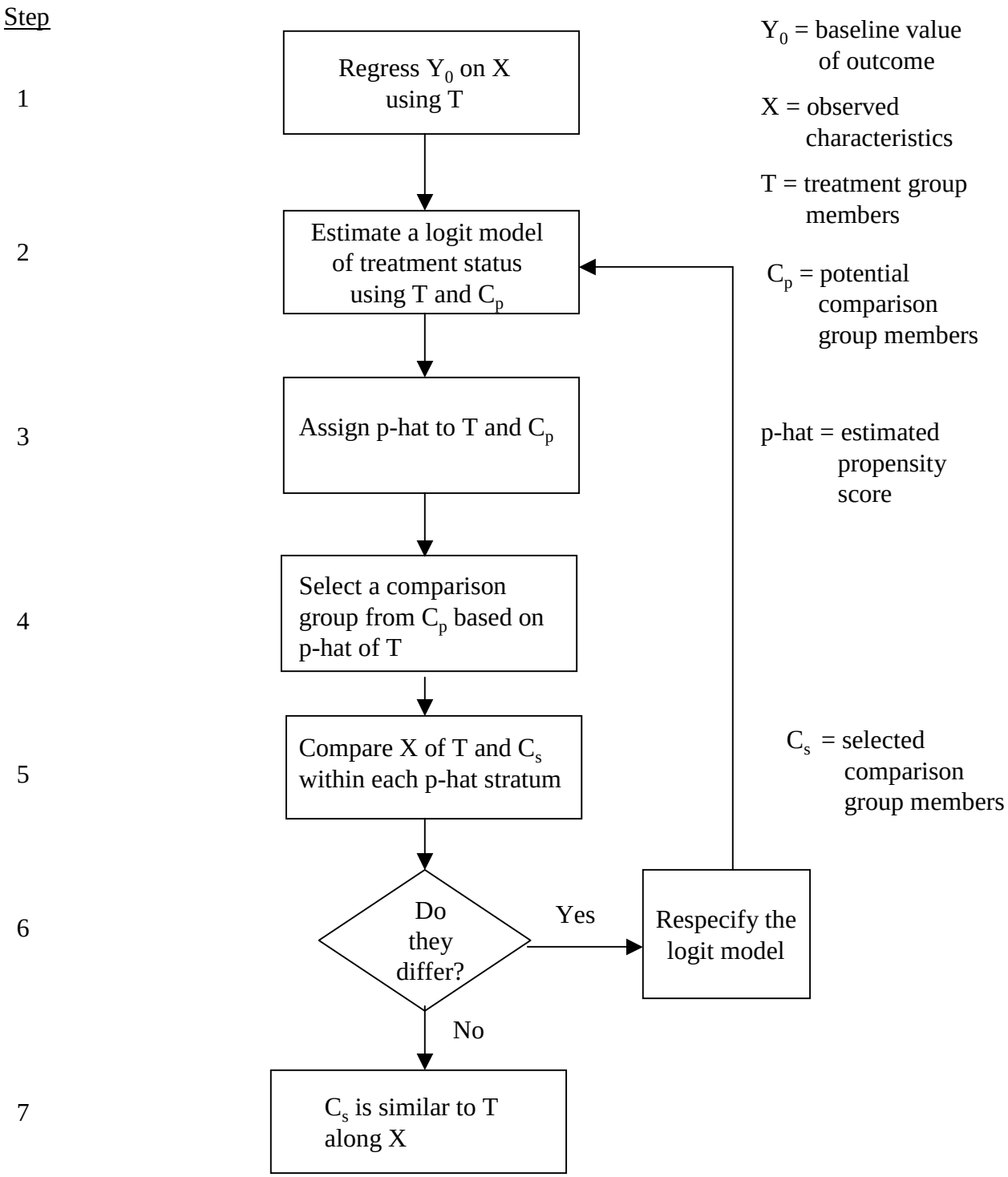

TABLE 2

SIMILARITY OF THE TREATMENT AND CONTROL/COMPARISON GROUPS
(P-value)[a]

| Program | Control Group | Comparison Group | |
|---|---|---|---|
| | | SDDAP | NELS |
| **Middle School** | | | |
| Albuquerque | 0.44 | 0.95 | 0.79 |
| Atlanta | 0.12 | 0.95 | 0.26 |
| Flint | 0.00 | 0.24 | 0.92 |
| Long Beach | 0.98 | 0.82 | 0.97 |
| Miami COMET | 0.75 | 0.41 | 0.92 |
| Newark | 0.01 | 0.24 | 0.16 |
| Rockford | 0.28 | 0.13 | 0.62 |
| Sweetwater | 0.24 | 0.12 | 0.92 |
| **High School** | | | |
| Boston | 0.00 | 0.32 | 0.07 |
| Chicago | 0.00 | 0.09 | 0.20 |
| Queens | 0.44 | 0.68 | 0.44 |
| Las Vegas | 0.12 | 0.38 | 0.06 |
| Miami Corp. Acad. | 0.95 | 0.25 | 0.94 |
| Seattle | 0.90 | 0.72 | 0.32 |
| St. Louis | 0.19 | 0.77 | 1.00 |
| Tulsa | 0.59 | 0.48 | 0.30 |

SOURCE: Authors' calculations based on data collected for the federal evaluation of the School Dropout Demonstration Assistance Program, and the National Education Longitudinal Study.

[a]Estimated p-value from F-test of the similarity of the collection of matching characteristics across the treatment and control/comparison groups.

represent the characteristics used in the matching process. A p-value of 0.05 or less indicates that the matching characteristics of a treatment and control/comparison group differ at conventional levels of statistical significance. Tables that report the similarity of the treatment and control/comparison groups along each matching characteristic are available upon request.[15]

These results indicate that the matching characteristics of the treatment and comparison groups are similar. In fact, for some treatment groups, their comparison groups are more similar than their control group along the matching characteristics. For example, the matching characteristics of the treatment and control groups for the Flint program are significantly different, whereas the treatment and comparison groups (both the SDDAP and NELS ones) for this program are not significantly different.

## V. METHODS

Experimental impacts were estimated as the difference in average outcomes between the treatment and the control group. Propensity score impacts were estimated in a similar way, using comparison groups instead of control groups. We also computed regression-adjusted propensity score impacts, as Rubin and Thomas (2000) suggest. In particular, using treatment and comparison group members, we regressed each outcome on a variable that indicates treatment status and variables that represent the characteristics used in the matching process. The regression-adjusted propensity score impact equals the coefficient on the treatment status variable. We computed these impacts to adjust for any differences that, while not statistically

[15]We also examined the similarity between each treatment group and a comparison group that included all SDDAP students, and each treatment group and a comparison group that included all NELS students. In all cases, the matching characteristics of treatment and these alternate comparison group members differ at the 0.01 level of confidence, two-tailed test.

significant, may nevertheless exist between the treatment and comparison groups. In our case, the simple and regression-adjusted impacts are similar. Therefore, we report only the simple impacts.

Standard errors for experimental impacts were computed using the standard analytic formula, whereas standard errors for propensity score impacts were computed using the bootstrap method (Efron 1982). We used the bootstrap method to compute standard errors for propensity score impacts, for two reasons. First, the treatment and comparison groups are not independent random samples, as the analytic formula assumes. Instead, the comparison group is selected based on the characteristics of the treatment group. This may reduce the standard error of propensity score impacts because it reduces by-chance differences between treatment and comparison groups, especially in small samples. Second, the criterion used to select the comparison group—the propensity score—is based on an estimate that may differ from its true value. Therefore, any difference between the estimated propensity score and its true value may increase the standard error of propensity score impacts.

Computing standard errors using the bootstrap method involves replicating the entire process used to compute propensity score impacts, which involves selecting a comparison group, many times. The accuracy with which the bootstrap approximates standard errors depends on the number of observations in the original sample and the number of times the bootstrap is repeated. In other work we have done, we found that bootstrap standard errors for complex statistics do not stabilize until the number of bootstrap replications approaches 1,000. This suggests that the entire process used to compute propensity score impacts should be replicated about 1,000 times. As such, it suggests that a comparison group must be selected for each of the bootstrap replicates.

Our bootstrap standard errors are based on 1,000 replications; however, to compute these standard errors more economically, we assumed that the specification of the logit model developed using the original sample was appropriate for each bootstrap sample (hereafter, "fixed logit model" standard errors). Specifically, we bootstrapped the entire process used to compute propensity score impacts, except for the step of respecifying the logit model in cases where the characteristics of treatment and comparison group members in a propensity score stratum differed. In terms of Figure 1, this means that we bootstrapped steps 2 through 7, but changed what happens at step 6 to counting the number of times the condition was met. To assess the appropriateness of this assumption, we counted the number of times the fixed logit model fit the bootstrap samples—that is, the number of bootstrap samples where the characteristics of treatment and comparison group members in each propensity score stratum were not statistically different. For comparison purposes, we also computed standard errors for propensity score impacts using the standard analytic formula (hereafter, "random sample" standard errors).

## VI. RESULTS

Before we present our results, it is important that we describe the criteria we used to determine whether propensity score methods replicate experimental impacts. In our setting, a program that had a negative impact is (what most would consider to be) an effective program because the program decreased a negative outcome. For example, a negative impact on the percent of students that dropped out means that the program increased the percent of students that graduated, or are attending high school. An effective program also is the type of program that policymakers are likely to use as a benchmark when deciding to fund other, similar programs. Therefore, we focus on the propensity score method's ability to detect programs that experimental methods indicate are effective. We also focus on situations in which experimental

methods indicate that a program was ineffective (that is, either had no effect or a positive impact), but propensity score methods indicate that the program was effective, as these are situations in which propensity score methods may lead policymakers to the wrong decision about whether to fund other similar programs.

## A. Impacts

Table 3 reports experimental and propensity score impacts on percent dropped out. The propensity score impacts include those based on both the SDDAP and NELS comparison groups. The proportion of treatment group members who dropped out is also reported. The results are reported separately for each of the dropout prevention programs grouped by middle-school and high-school program.

These results do not provide consistent evidence that propensity score methods replicate experimental impacts of dropout prevention programs. The experimental results indicate that 3 of the 16 dropout prevention programs—Atlanta, Flint, and Miami COMET—were effective at reducing dropping out. Had these programs been evaluated using propensity score methods based on the SDDAP comparison groups, the effectiveness of the Atlanta program would have been detected. The experimental impact indicates that the Atlanta program reduced dropping out by 11.4 percentage points, whereas the SDDAP propensity score impact indicates that the reduction was 14.4 percentage points. However, the effectiveness of the Flint and Miami COMET programs would not have been detected. The experimental impacts indicate that the Flint and Miami COMET programs reduced dropping out by 9.5 and 5.0 percentage points, respectively, whereas the SDDAP propensity score impacts indicate that these programs reduced dropping out by 4.5 and 11.7 percentage points, respectively. However, neither of these SDDAP propensity score impacts are statistically significant. Inferences about the effectiveness of one of

TABLE 3

IMPACTS ON PERCENT DROPPED OUT

| Program | Treatment Group Dropout Rate | Impact | | |
|---|---|---|---|---|
| | | | Propensity Score | |
| | | Experimental | SDDAP | NELS |
| **Middle School** | | | | |
| Albuquerque | 11.5 | 1.8 | 5.5 | 4.1 |
| Atlanta | 4.6 | -11.4* | -14.4* | 0.8 |
| Flint | 1.0 | -9.5* | -4.5 | -4.4 |
| Long Beach | 4.7 | -0.1 | -2.9 | -0.5 |
| Miami COMET | 0.0 | -5.0* | -11.7 | -3.4 |
| Newark | 6.1 | 0.7 | -11.7* | 4.5 |
| Rockford | 6.5 | -1.0 | -3.2 | 0.2 |
| Sweetwater | 7.4 | 0.3 | 2.2 | 3.9 |
| **High School** | | | | |
| Boston | 32.6 | 2.7 | 10.2 | -12.2 |
| Chicago | 12.6 | 5.9 | 1.6 | 4.7 |
| Queens | 39.5 | -7.4 | 9.7 | -32.9* |
| Las Vegas | 54.5 | 7.6 | 29.4* | 28.7* |
| Miami Corp. Acad. | 31.8 | -2.1 | -2.2 | -31.9* |
| Seattle | 35.0 | 1.3 | 18.3* | -28.1* |
| St. Louis | 61.8 | -1.4 | 44.9* | 50.2* |
| Tulsa | 65.7 | 1.0 | 43.8* | -14.0* |

SOURCE: Authors' calculations based on data collected for the federal evaluation of the School Dropout Demonstration Assistance Program, and the National Education Longitudinal Study.

*Significantly different from zero at the 0.05 level, two-tailed test.

the other 13 programs—the Newark program—also would have been different, as the experimental impacts indicate that it was ineffective, whereas the SDDAP propensity score impacts indicate that it was effective.

Results based on the NELS comparison groups provide more evidence that propensity score methods do not replicate experimental impacts in our settings. The NELS propensity score impacts indicate that the 3 programs that experimental impacts indicate were effective at reducing dropping out, did not have an effect. Also, the NELS propensity score impacts indicate that several of the other 13 programs were effective, whereas the experimental impacts indicate that none of them were effective. SDDAP and NELS propensity score impacts for the three other outcomes we analyzed are consistent with these findings and are reported in Tables A.3, A.4, and A.5 in the appendix.

We also find no patterns in our results that suggest the situations in which propensity score methods replicate experimental impacts. Rosenbaum (1987) showed that, if propensity score impacts based on comparison groups selected from different data sources are similar, important unobserved characteristics are likely to have been captured by the matching process. As a result, propensity and experimental impacts should be similar. This is not the case in our setting. For example, the SDDAP and NELS propensity score impacts on dropping out for the Flint and Las Vegas programs are similar. However, they are considerably different than the experimental impacts for these programs. In addition, the success of the matching process does not, in our setting, suggest situations in which propensity score methods replicate experimental impacts. For example, the SDDAP comparison group for the Flint treatment group was selected using 17 of the 20 characteristics we wanted to use in the matching process, whereas the SDDAP comparison group for the Atlanta treatment group was selected using 15 characteristics. Moreover, the two additional characteristics—reading and math test scores—that were used to

select the SDDAP comparison group for the Flint treatment group are characteristics that many evaluators would consider important. Nevertheless, propensity score methods did not detect the effectiveness of the Flint program at reducing dropping out, but did detect the effectiveness of the Atlanta program.

These findings suggest that, even though the data we used to select comparison groups are extensive by most standards, our comparison groups differ from their respective treatment groups in important ways that we do not observe. As mentioned above, a key assumption of propensity score methods is that all characteristics related to treatment status that are also related to outcomes must be observed. The SDDAP and NELS comparison groups may not satisfy this condition because we do not observe, for example, the extent to which students were motivated to succeed in school and therefore were interested in receiving program services. Our results suggest that our comparison groups differ from their respective treatment groups in these ways.

An interesting issue is whether other methods that are easier to implement than propensity score methods are equally capable, or more capable of replicating experimental impacts, than propensity score methods. We explored this issue by estimating impacts using regression models. In particular, using treatment and all potential comparison group members, we regressed dropout status on treatment status and the characteristics used in the matching process. Table 4 reports the coefficient on the treatment status variable.

The results indicate that regression-based impacts are not more capable than propensity score methods of replicating experimental impacts. For example, the SDDAP regression-based impacts detect the effectiveness of the Flint program, but not the effectiveness of the Atlanta program. The opposite is true of the SDDAP propensity score impacts. Also, like the propensity score impacts, the regression-based impacts indicate that several of the other programs had an effect on dropping out and often reduced it, whereas the experimental impacts indicate that none

TABLE 4

REGRESSION-BASED IMPACTS ON PERCENT DROPPED OUT

| Program | Experimental | Regression-Based | |
|---|---|---|---|
| | | All SDDAP Students | All NELS Students |
| **Middle School** | | | |
| Albuquerque | 1.8 | 3.0 | 1.6 |
| Atlanta | -11.4* | -6.4 | -3.7 |
| Flint | -9.5* | -10.6* | -5.9* |
| Long Beach | -0.1 | 0.5 | 0.9 |
| Miami COMET | -5.0* | -12.3* | -4.8 |
| Newark | 0.7 | -5.1* | -0.6 |
| Rockford | -1.0 | -1.7 | 1.1 |
| Sweetwater | 0.3 | 0.5 | 3.5* |
| **High School** | | | |
| Boston | 2.7 | 10.4* | -2.0 |
| Chicago | 5.9 | 3.5 | 0.2 |
| Queens | -7.4 | 16.7* | -37.7* |
| Las Vegas | 7.6 | 34.7* | 31.9* |
| Miami Corp. Acad. | -2.0 | 0.9 | -21.2* |
| Seattle | 1.3 | 11.8* | -27.2* |
| St. Louis | -1.4 | 40.3* | 37.3* |
| Tulsa | 1.0 | 42.1* | 0.1 |

SOURCE: Authors' calculations based on data collected for the federal evaluation of the School Dropout Demonstration Assistance Program, and the National Education Longitudinal Study.

*Significantly different from zero at the 0.05 level, two-tailed test.

of them had an effect. These results provide more support for the notion that important characteristics have been excluded from the matching process.

### B. Standard Errors

Table 5 reports standard errors of the experimental and propensity score impacts on percent dropped out. Standard errors of the propensity score impacts include those based on the SDDAP and NELS comparison groups. Also, standard errors of the propensity score impacts include the two we described earlier: (1) fixed logit model, and (2) random sample. The fixed logit model standard errors were computed using bootstrap methods and assume that the specification of the logit model developed using the original sample was appropriate for each of the 1,000 bootstrap samples. The random sample standard errors assume that, taken together, the treatment and comparison groups are a random sample and use the standard analytic formula. The table also reports the situations in which the fixed logit model assumption was appropriate for at least 90 percent of the bootstrap samples.

These results indicate that the way in which standard errors of propensity score impacts are computed does not seem to matter much. Standard errors based on the fixed logit model and random sample assumptions are similar. In fact, some of the standard errors for the propensity score impacts are smaller than those for the experimental impacts. For example, the standard errors of the NELS propensity score impacts for the Atlanta program (4.4 for the fixed logit model ones and 4.2 for the random sample ones) are smaller than the experimental one (5.4).[16] This is consistent with Rubin and Thomas (1992) who showed that matching can reduce variance, especially in small samples.

[16]The Atlanta program is a situation in which the fixed logit model assumption was appropriate for at least 90 percent of the SDDAP and NELS bootstrap samples.

TABLE 5

STANDARD ERRORS FOR IMPACTS ON PERCENT DROPPED OUT

| Program | Experimental | Propensity Score: SDDAP Fixed Logit Model | Propensity Score: SDDAP Random Sample | Propensity Score: NELS Fixed Logit Model | Propensity Score: NELS Random Sample |
|---|---|---|---|---|---|
| **Middle School** | | | | | |
| Albuquerque | 3.9 | 4.2[a] | 3.4 | 4.3 | 4.4 |
| Atlanta | 5.4 | 6.0[a] | 6.2 | 4.4[a] | 4.2 |
| Flint | 3.5 | 5.1 | 3.9 | 3.0[a] | 3.2 |
| Long Beach | 2.8 | 2.9[a] | 3.1 | 4.0[a] | 3.6 |
| Miami COMET | 2.3 | 8.8 | 4.6 | 3.4[a] | 2.2 |
| Newark | 2.4 | 6.0 | 4.7 | 3.9 | 4.2 |
| Rockford | 2.4 | 4.4 | 3.1 | 2.4[a] | 3.4 |
| Sweetwater | 2.5 | 3.3[a] | 2.7 | 2.3[a] | 3.2 |
| **High School** | | | | | |
| Boston | 6.7 | 13.2 | 13.3 | 13.0 | 13.0 |
| Chicago | 4.6 | 6.2[a] | 5.5 | 7.7 | 7.1 |
| Queens | 9.5 | 11.5[a] | 10.1 | 15.2 | 15.0 |
| Las Vegas | 5.3 | 7.8 | 6.9 | 6.5 | 6.5 |
| Miami Corp. Acad. | 8.6 | 25.5 | 19.1 | 13.0[a] | 11.4 |
| Seattle | 5.0 | 7.1 | 7.0 | 9.9 | 10.5 |
| St. Louis | 6.6 | 8.0 | 9.6 | 8.8 | 16.5 |
| Tulsa | 5.5 | 10.4 | 12.7 | 5.0 | 10.3 |

SOURCE: Authors' calculations based on data collected for the federal evaluation of the School Dropout Demonstration Assistance Program, and the National Education Longitudinal Study.

[a]Fixed logit model was appropriate for at least 90 percent of the bootstrap samples.

# VII. CONCLUSIONS

We find no consistent evidence that propensity score methods replicate experimental impacts of the dropout prevention programs funded by the SDDAP. In fact, we find that evaluating these programs using propensity score methods might have led to misleading inferences about the effectiveness of the programs. This is true for propensity score impacts based on extensive data and those based on less extensive data that researchers are likely to have access to. We also find that impacts based on regression methods, which are easier to implement, are not any more capable of replicating experimental impacts in this setting than are propensity score methods.

The theoretical basis for propensity score methods rests on the assumption that all the characteristics related to treatment status, that are also related to outcomes, are observed. This also is the case for many other nonexperimental methods, such as regression methods. The SDDAP programs targeted students who were at risk of dropping out. Our propensity score impacts are based on extensive data that are not typically available to researchers, and that contain information that can be used to determine the types of students that were targeted by the programs. However, whether even these data contain enough information to capture the extent to which students were interested in receiving program services—the criteria that, in addition to being at risk, determined the types of students served by the programs—is an open question. Our results suggest that even these data, though extensive by most standards, do not contain enough information to produce reliable propensity score impacts in our setting.

These findings suggest that evaluators need to consider carefully the process by which programs target individuals and the process by which individuals enter programs. In situations where individuals enter programs of their own volition, unobserved factors such as motivation may be exerting powerful influences on outcomes, influences not easily captured using

nonexperimental methods. It would be useful to explore the propensity score method's ability to replicate experimental results in settings where participation is more directed or mandatory, in which case, unobservable factors may be less influential.

## REFERENCES


Ashenfelter, Orley. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics*, vol. 60, 1978, pp. 47-57.

Ashenfelter, Orley, and David Card. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics*, vol. 67, 1985, pp. 648-660.

Dehejia, Rajeev H., and Sadek Wahba. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*, vol. 94, no. 448, 1999, pp. 1053-1062.

Dehejia, Rajeev H., and Sadek Wahba. "Propensity Score Matching Methods for Nonexperimental Causal Studies." NBER Working Paper No. 6829, 1998.

Efron, Bradley. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics, 1982.

Fraker, Thomas, and Rebecca Maynard. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources*, vol. 22, no. 2, 1987, pp. 194-227.

Friedlander, Daniel, and Philip K. Robins. "Estimating the Effects of Employment and Training Programs: An Assessment of Some Nonexperimental Techniques." Working Paper. New York: MDRC, 1995.

Gleason, Philip, and Mark Dynarski. "Falling Behind: Characteristics of Students in Federally Funded Dropout Prevention Programs—Part Two: Restructuring Projects." Princeton, NJ: Mathematica Policy Research, Inc., March 1995.

Gleason, Philip, and Mark Dynarski. "Falling Behind: Characteristics of Students in Federally Funded Dropout Prevention Programs—Part One: Targeted Projects." Princeton, NJ: Mathematica Policy Research, Inc., September 1994.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. "Characterizing Selection Bias Using Experimental Data." *Econometrica*, vol. 66, no. 5, 1998, pp. 1017-1098.

Lalonde, Robert. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, vol. 76, no. 4, 1986, pp. 604-620.

Rosenbaum, Paul R. "The Role of a Second Control Group in an Observational Study." *Statistical Science*, vol. 2, 1987, pp. 292-316.

Rosenbaum, Paul R. *Observational Studies.* New York: Springer-Verlag, 1995.

Rosenbaum, Paul R., and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, vol. 70, 1983, pp. 41-55.

Rubin, Donald B. "Matching to Remove Bias in Observational Studies." *Biometrics*, vol. 29, 1973, pp. 159-183.

Rubin, Donald B. "Use of Propensity Scores for Tobacco Litigation." Working Paper. Cambridge, MA: Harvard University, 2001.

Rubin, Donald B., and Neal Thomas. "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates." *Journal of the American Statistical Association*, vol. 95, no. 450, 2000, pp. 573-585.

Rubin, Donald B., and Neal Thomas. "Characterizing the Effect of Matching Using Linear Propensity Score Method's with Normal Distributions." *Biometrika*, vol. 79, no. 4, 1992, pp. 797-809.

Smith, Jeffrey, and Petra Todd. "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" Working Paper on website http://www.ssc.uwo.ca/economics/faculty/JSmith/nsw112200.pdf. Accessed January 2001.

Spencer, B. D., M. R. Frankel, S. J. Ingels, K. A. Rasinski, and R. Tourangeau. *NELS:88 Base Year Sample Design Report.* Washington, DC: National Center for Education Statistics, 1990.

# APPENDIX

## CHARACTERISTICS USED TO SELECT COMPARISON GROUPS AND IMPACTS ON OTHER OUTCOMES

TABLE A.1

CHARACTERISTICS USED TO SELECT SDDAP COMPARISON GROUPS

| | Program | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Middle School | | | | | | | | High School | | | | | | | |
| | Albuquerque | Atlanta | Flint | Long Beach | Miami Comet | Newark | Rockford | Sweetwater | Boston | Chicago | Queens | Las Vegas | Miami Corp. Acad. | Seattle | St. Louis | Tulsa |
| **Baseline Outcome** | | | | | | | | | | | | | | | | |
| Dropout | NA | NA | NA | NA | NA | NA | NA | NA | X | ✓ | X | X | ✓ | X | X | X |
| Absenteeism | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | NA | ✓ | NA | ✓ | ✓ | NA | NA | NA |
| Educ. expectations | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Self-esteem | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **Student Characteristics** | | | | | | | | | | | | | | | | |
| Age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Sex | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Race/ethnicity | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Reading score | ✓ | NA | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | NA | ✓ | NA | ✓ | ✓ | NA | NA | NA |
| Math score | NA | NA | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | NA | ✓ | NA | ✓ | ✓ | NA | NA | NA |
| Fun reading | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Mother's educ. | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | X | ✓ | X | ✓ | ✓ | X | ✓ | ✓ |
| Ever dropout | NA | NA | NA | NA | NA | NA | NA | NA | X | X | X | X | ✓ | X | X | X |
| Homework time | ✓ | ✓ | ✓ | X | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X |
| Sibling dropout | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Extracurricular | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Both parents | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Overage for grade | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Discipline problems | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Locus of control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **School Characteristics** | | | | | | | | | | | | | | | | |
| Urbanicity | X | X | X | ✓ | ✓ | ✓ | X | X | ✓ | ✓ | ✓ | X | ✓ | ✓ | X | X |

✓ = Used in the matching process.

NA = Not used in the matching process because not available.

X = Available but could not be used in the matching process.

TABLE A.2

CHARACTERISTICS USED TO SELECT NELS COMPARISON GROUPS

| | Program | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Middle School | | | | | | | | High School | | | | | | | |
| | Albuquerque | Atlanta | Flint | Long Beach | Miami Comet | Newark | Rockford | Sweetwater | Boston | Chicago | Queens | Las Vegas | Miami Corp. Acad. | Seattle | St. Louis | Tulsa |
| **Baseline Outcome** | | | | | | | | | | | | | | | | |
| Dropout | NA | NA | NA | NA | NA | NA | NA | NA | X | ✓ | X | X | ✓ | X | X | X |
| Educ. expectations | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | X | ✓ | X |
| Self-esteem | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ |
| **Student Characteristics** | | | | | | | | | | | | | | | | |
| Age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Sex | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Race/ethnicity | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Mother's educ. | ✓ | ✓ | X | X | X | X | X | X | X | X | X | ✓ | ✓ | X | X | X |
| Ever dropout | NA | NA | NA | NA | NA | NA | NA | NA | X | ✓ | ✓ | X | ✓ | ✓ | X | ✓ |
| Homework time | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | X |
| Both parents | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Overage for grade | ✓ | X | X | X | X | X | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Locus of control | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ |
| **School Characteristics** | | | | | | | | | | | | | | | | |
| Urbanicity | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

✓ = Used in the matching process.

NA = Not used in the matching process because not available.

X = Available but could not be used in the matching process.

TABLE A.3

IMPACTS ON PERCENT OF DAYS ABSENT

| Program | Treatment Group Mean | Impact | | |
|---|---|---|---|---|
| | | | Propensity Score | |
| | | Experimental | SDDAP | NELS |
| **Middle School** | | | | |
| Albuquerque | 8.2 | 2.2 | -4.2 | -- |
| Atlanta | 17.6 | 2.4 | -0.9 | -- |
| Flint | 12.7 | -5.1* | -2.0 | -- |
| Long Beach | 9.4 | -0.5 | 3.0 | -- |
| Miami COMET | 6.6 | 1.1 | -2.9 | -- |
| Newark | 14.9 | 3.5* | 5.8* | -- |
| Rockford | 17.6 | -1.2 | 3.6 | -- |
| Sweetwater | 5.7 | -0.3 | -5.9* | -- |
| **High School** | | | | |
| Boston | -- | -- | -- | -- |
| Chicago | 16.3 | 1.2 | 4.4 | -- |
| Queens | -- | -- | -- | -- |
| Las Vegas | 21.5 | -1.5 | -2.5 | -- |
| Miami Corp. Acad. | 17.6 | -5.0 | 2.0 | -- |
| Seattle | -- | -- | -- | -- |
| St. Louis | -- | -- | -- | -- |
| Tulsa | -- | -- | -- | -- |

SOURCE: Authors' calculations based on data collected for the federal evaluation of the School Dropout Demonstration Assistance Program, and the National Education Longitudinal Study.

-- Days absent not available.

*Significantly different from zero at the 0.05 level, two-tailed test.

TABLE A.4

IMPACTS ON PERCENT EXPECTING TO COMPLETE HIGH SCHOOL OR LESS

| Program | Treatment Group Mean | Impact | | |
|---|---|---|---|---|
| | | | Propensity Score | |
| | | Experimental | SDDAP | NELS |
| **Middle School** | | | | |
| Albuquerque | 18.9 | 5.2 | 0.0 | -28.5* |
| Atlanta | 26.9 | -15.7 | 7.7 | -18.6 |
| Flint | 13.0 | -1.4 | -18.8 | -25.2* |
| Long Beach | 14.1 | 0.4 | -5.3 | -14.3 |
| Miami COMET | 5.2 | 5.2 | -4.2 | -21.2* |
| Newark | 8.2 | 1.3 | -0.8 | -14.3 |
| Rockford | 24.2 | 2.8 | -1.1 | -15.2* |
| Sweetwater | 4.9 | 1.4 | -4.0 | -24.5* |
| **High School** | | | | |
| Boston | 11.5 | 2.6 | -0.6 | -46.9* |
| Chicago | 18.0 | -1.0 | -4.8 | -26.8* |
| Queens | 18.8 | -7.9 | 3.6 | -74.7* |
| Las Vegas | 34.6 | 11.7* | 1.9 | -29.2* |
| Miami Corp. Acad. | 18.0 | -5.4 | -21.0 | -60.3* |
| Seattle | 9.5 | -9.0* | -2.3 | -83.4* |
| St. Louis | 32.2 | 6.8 | 11.1 | -11.2 |
| Tulsa | 19.0 | -4.8 | -2.7 | -46.5* |

SOURCE: Authors' calculations based on data collected for the federal evaluation of the School Dropout Demonstration Assistance Program, and the National Education Longitudinal Study.

*Significantly different from zero at the 0.05 level, two-tailed test.

TABLE A.5

IMPACTS ON PERCENT WITH LOW TO MODERATE SELF-ESTEEM

| Program | Treatment Group Mean | Impact | | |
|---|---|---|---|---|
| | | | Propensity Score | |
| | | Experimental | SDDAP | NELS |
| **Middle School** | | | | |
| Albuquerque | 78.3 | -4.5 | 1.1 | 7.1 |
| Atlanta | 50.8 | -16.6 | -9.3 | -33.3* |
| Flint | 76.0 | 10.0 | 7.5 | 3.9 |
| Long Beach | 73.3 | -5.6 | 0.0 | -3.0 |
| Miami COMET | 73.2 | -4.0 | -3.2 | -11.4 |
| Newark | 55.8 | 6.7 | -4.8 | -10.4 |
| Rockford | 67.4 | -2.8 | -3.6 | -9.9 |
| Sweetwater | 67.9 | 1.3 | 3.2 | -3.9 |
| **High School** | | | | |
| Boston | 58.3 | 0.5 | 6.8 | -24.9 |
| Chicago | 64.9 | -1.9 | -6.4 | 9.0 |
| Queens | 70.4 | -3.8 | 2.9 | 1.8 |
| Las Vegas | 61.1 | -7.6 | -12.9* | -15.9* |
| Miami Corp. Acad. | 54.0 | -16.0 | -22.3 | -16.5 |
| Seattle | 58.0 | -1.5 | -0.6 | -5.1 |
| St. Louis | 64.3 | 3.8 | 4.4 | -15.8 |
| Tulsa | 63.9 | -2.9 | -11.0 | 18.2 |

SOURCE: Authors' calculations based on data collected for the federal evaluation of the School Dropout Demonstration Assistance Program, and the National Education Longitudinal Study.

*Significantly different from zero at the 0.05 level, two-tailed test.