Title: Two Approaches to Quasi-experimental Program Evaluation Using State-wide Educational Data Systems: Results of Computational Experiments

Authors and Affiliations:

Valeriy Lazarev, Empirical Education Inc., vlazarev@empiricaleducation.com
Andrew Jaciw, Empirical Education Inc., ajaciw@empiricaleducation.com
Denis Newman, Empirical Education Inc., dn@empiricaleducation.com
Abstract Body

Background:

The practical importance of quasi-experimental (QE) studies in education has been growing due to increasing availability and reliability of longitudinal databases maintained by state education agencies. These databases present special challenges and opportunities because of their size and comprehensive nature. With access to such a database, a program implemented anywhere in a state (for example, in a few schools in one district), can be evaluated by drawing comparable units from other districts across the state. However, program evaluation that uses observational data is affected by the non-equivalence of treated and untreated units, which leads to biased estimates of program effect. A variety of analytical strategies have been developed to approximate experimental conditions and therefore minimize the selection bias. Though a considerable body of research has been devoted to the analysis of empirical problems of QE studies of job training programs, very little work of this kind has been conducted in the context of large-scale evaluation in K-12 education (see Bloom, Michalopoulos, and Hill (2005) and Glazerman, Levy, and Myers (2003) for recent contributions and surveys of earlier studies). In addition, existing program evaluation methods have been developed and tested on relatively small datasets—typically thousands of observations—while a student information system of a typical state or even a large district contains millions of records. Given the limited computing resources available to most mainstream and many academic researchers, the issues of computational efficiency of analytical methods come to the foreground.

Purpose and Focus of Study:

In this paper, we present the results of computational experiments that compare statistical performance and computational efficiency of two analytical methods in simulated QE studies. The main goal of these computational experiments is to test the statistical properties and performance of two alternative approaches to QE estimation of program impact using a large student database: 1) two-stage multivariate matching (2MM) and 2) a widely-used alternative - propensity score matching (PSM). The first method assumes finding best matches at the school level to the treatment schools and student-level matching within the clusters of matched schools identified in the first stage. The second method relies on calculating a single propensity function using the entirety of observations and both school and student-level variables, followed by nearest-neighbor matching. The major difference between the two methods is that 2MM promises higher computational efficiency because of its reliance on a much smaller number of observations while PSM can be expected to produce more accurate estimates because it utilizes more information and relies on a more complex selection model. The issue of the potential tradeoff between the computational efficiency and statistical quality is of primary interest for this study.

Setting:

N/A

Population:
The dataset used in these computational experiments included 4th through 8th grade public school students within one state in the U.S.

**Intervention / Program / Practice:**

N/A

**Research Design:**

We use a version of the methodology proposed first by Lalonde (1986) to simulate a quasi-experimental study and estimate the quality of its results using a dataset and benchmark estimates from an RCT. We use a dataset produced as a result of a large multi-site RCT where treatment and control schools were matched on a broad variety of school and student-level variables obtained in part through a survey. The original controls are then removed, small random samples of treatment schools are drawn from the dataset, and best matches are found using the two methods to be compared. Only a small set of variables that are likely to be found in state student databases are used in the analysis. The known (positive and significant) RCT estimate of the program effect is used as the benchmark.

To imitate a large-scale QE study involving a small number of pilot program schools evaluated against a state student performance metric, each iteration of the experiment started with drawing a small random sample of treatment schools (four schools, or approximately 10% of the total). Original controls that were paired with the selected treatment schools and remaining treatment schools were removed from the dataset. Then for 2MM, in the first step, each treatment school in the sample was matched (without replacement) to five of the remaining schools with the lowest Mahalanobis distance calculated on the basis of school characteristics and school-average student characteristics. Then students were matched with replacements within clusters of matched schools, also using the Mahalanobis metric.

An alternative analysis was performed using nearest-neighbor propensity score matching with regression adjustment (Dehejia and Wahba, 2002; Rubin and Thomas, 2000). The propensity function was calculated using all available variables and all treatment and remaining control schools. Propensity score matching was performed at the student level within the entire dataset.

**Statistical Model:**

For the purposes of program evaluation using a state database, which involves a relatively small number of treatment units and a very large number of potential comparison units, it is reasonable to adopt a two-stage analytical strategy, separating matching at the school level between participating and comparison schools from student-level analyses. Two groups of arguments in favor of this strategy can be presented: fundamental and technical.

First, in a real-world program adoption (not in the framework of a large-scale RCT), it is typically the case that schools or districts are the agents who make the program adoption decisions. The decision of the program developer to select a school and/or the decision of a
school to self-select into treatment is contingent on the school-level variables. The subsequent decision to select students into treatment is made on the basis of individual student characteristics (and possibly class averages) within the constraints imposed by the first stage selection. Schematically, the relationship that drives the selection of students into treatment in such a process can be considered additively-separable:

\[ T_{ij} = F(X_{ij}) = F_1(X_j) + F_2(X_{ij}') \]

where covariates \( X_{ij} \) (\( i=\)student, \( j=\)school) are composed of school means \( X_j \) and the deviations from the school mean \( X_{ij}' = X_{ij} - X_j \). In this case, matching can be performed in two stages - first on the school level, then on the student level within schools - without loss of accuracy.

Second, more generally, the proximal cause of student progress, as Raudenbush (2008) argues is the “instructional regime” – a combination of observable and unobservable factors that operate at the school level. It therefore makes sense to single out school-level matching. Furthermore, factors driving program adoption in a multi-site setting can differ across schools and the differences in the propensity to participate in a program are of a research interest per se. This calls for estimating separate propensity functions at the school and student levels (Hong & Raudenbush, 2006).

Two-stage matching may also be preferred for the following technical reasons. First, a typical state database will contain 1M – 10M student records. One-stage matching using all student data from one state may not be feasible for reasons of computing capacity. Some sort of two-stage solution will be required anyway, e.g. random sampling or sequential processing.

Second, two-stage matching provides a more flexible solution in terms of analysis because it allows using different analytical approaches at the school and student levels. This may be particularly important when selection into treatment at the school level is contingent on school-level variables \( Z_j \) which are not averages of student characteristics and have statistical properties that are not found among the student-level variables (e.g. expenditure per student tends to be heavily skewed). Although we used the same procedure – multivariate matching with Mahalanobis weighting – at both stages, this is not required under the proposed approach.

**Usefulness / Applicability of Method:**

The proposed 2MM approach outperforms the alternative method both in terms of statistical quality and computational efficiency. See ‘Results’ section below.

**Data Collection and Analysis:**

As a basis for testing, we used a dataset produced as a result of an experimental study of a new STEM program for elementary and middle school students. The study was a cluster randomized trial with close to one hundred schools assigned in matched pairs to treatment or control. In total, the study produced close to 100K student records spanning several years. An array of student demographic variables (gender, ethnicity, free lunch status, and English language proficiency) and school-level covariates (location/community type and expenditure per student) were available. A subset that included approximately 20K students with complete data for the first
year of program implementation and one year prior was used to produce the estimate of program impact on student achievement in math as measured by state test scores.

Findings / Results:

The main body of our study consisted of 300 computational experiments (which corresponds to approximately .25% of all possible school combinations in the dataset) using the original dataset with ~20K student records to analyze the statistical properties of the estimates produced by the two methods and their relative speed. In both cases, results were adjusted using OLS and exact matching on the grade level was used for comparability with the original estimate, which was moderated by the grade level. The results are shown in Table 1.

In addition, we performed experiments aimed at establishing the relationship between the sample size and the computational performance of PSM. Computational performance (average time to complete) of both 2MM and PSM depends on the total number of schools in the state database, the number of treatment schools, and the average number of students per school. However 2MM’s dependence on the total number of schools is trivial as it affects only the first school-level stage of estimation, where the size of the dataset remains on the order of ~1K. PSM as implemented in this study, however, would depend on the total number of students in the state database. Therefore, it is critical to evaluate the completion time as a function of the total dataset size. It can be expected that an efficiently implemented PSM would have the computational complexity of O(n log(n)) – the same as an efficient sorting algorithm. This means that the computational time would increase rapidly at the small sample size, becoming nearly linear as the sample size increases. The results of the test that was performed by simple replication of records in the original dataset were consistent with this prediction. These results are presented in Figure 1.

Conclusions:

This study shows that PSM is hardly feasible when using standard hardware-software bundles to analyze datasets of a size typical of medium to large state student information systems. It is no surprise that 2MM, designed with the goal of computational efficiency, outperforms PSM computationally at all data sizes, although this advantage is of low importance when the datasets are relatively small (~10K-100K records). It comes as a surprise, however, that it outperformed PSM by far in the accuracy of the estimates. One possible explanation is that the experimental treatment group and the simulated QE control groups remained relatively well balanced despite the disturbances that we introduced into the derivative datasets. While further experiments focused on modeling stronger selection biases in the evaluation data are needed, current results suggest that the two-stage is a viable approach that is capable of yielding valid program impact estimates with large datasets while requiring reasonable computer resources. Moreover, it is possible that it is the unobserved variability in local conditions that results in inferior performance of PSM, suggesting that 2MM may be a preferred approach to deal with “big data” in educational program and policy evaluation.
Appendix A. References


Appendix B. Tables and Figures

Table 1. Comparative performance of 2MM and PSM methods.

<table>
<thead>
<tr>
<th></th>
<th>2MM</th>
<th>PSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average effect estimate</td>
<td>1.79</td>
<td>3.97</td>
</tr>
<tr>
<td>Average bias</td>
<td>0.10</td>
<td>2.28</td>
</tr>
<tr>
<td>Average bias, % of RCT estimate</td>
<td>5</td>
<td>135</td>
</tr>
<tr>
<td>St.dev. of average estimate</td>
<td>5.19</td>
<td>3.54</td>
</tr>
<tr>
<td>Average std. error of the estimate</td>
<td>1.69</td>
<td>3.51</td>
</tr>
<tr>
<td>Proportion of estimates with p &lt;0.05, %</td>
<td>26</td>
<td>2</td>
</tr>
<tr>
<td>Average time to complete, sec.</td>
<td>1.6</td>
<td>58.0</td>
</tr>
</tbody>
</table>

Notes: 1) Tests score scale, 2) Difference between the average estimate and the benchmark RCT estimate (1.69), 3) ~20,000 records

Figure 1. Computer time needed to produce PSM estimates.