Abstract Title Page
Not included in page count.

Title: Comparison of Student-Level and School-Level Data in a National Impact Evaluation

Authors and Affiliations:
Velez, Melissa, Abt Associates, Inc., Melissa_Velez@abtassoc.com
Sahni, Sarah, Abt Associates, Inc., Sarah_Sahni@abtassoc.com
Rulf-Fountain, Alyssa, Abt Associates, Inc., Alyssa_Rulf_Fountain@abtassoc.com
Gamse, Beth, Abt Associates Inc., Beth_Gamse@abtassoc.com

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

The ability to understand what works in education requires rigorous research methodologies with which to make informed, data-driven decisions. Increasingly, only the most promising education interventions with evidence based on rigorous research are receiving federal dollars. Demonstrating this level of efficacy often requires impact evaluations, which produce some of the highest standards of evidence. Unfortunately, collecting the individual-level data necessary for these kinds of evaluations is increasingly becoming a barrier to the research.

One of the primary obstacles facing education researchers today is the struggle to obtain student-level data from states, districts, and schools. Researchers typically face one of two scenarios; they must either 1) work with contractors hired by the state or district to handle data requests who can be prohibitively expensive or 2) invest significant amounts of time making data requests from overworked school district staff that can take a year or longer. Due to budget cuts and waning resources, district staff juggle many competing demands and typically prioritize district-level needs above those of external evaluators. It is becoming increasingly clear the research community must look for methodological alternatives that allow the continuation of rigorous research about what works in education.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

We explore the possibility of using publicly available school-level data instead of student-level data to conduct an impact analysis of a national education intervention. We use a national evaluation of an Expanded Learning Time (ELT) initiative as a case-study for these methods.

Setting:

Description of the research location.
(May not be applicable for Methods submissions)

Middle schools located across the country.

Population / Participants / Subjects:

Description of the participants in the study: who, how many, key features, or characteristics.
(May not be applicable for Methods submissions)

The treatment group consisted of eight ELT schools that implemented the initiative in the 2010-11 school year. Each ELT school was matched with up to four comparison schools with similar demographic and achievement profiles within the same district. The matching process used the Mahalanobis distance matching method (Rubin, 1980 and 1997) with school-level extant data from up to five years prior to ELT implementation on student demographic characteristics, test scores, and other school characteristics for which data were available.

Intervention / Program / Practice:

Description of the intervention, program, or practice, including details of administration and duration.

(May not be applicable for Methods submissions)

The ELT initiative mobilizes a second shift of educators, who provide academic support, leadership development, and "apprenticeships" after the end of the traditional school day. Students in the ELT schools are required to participate in the ELT programming.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

It is becoming increasingly evident that the education research community may no longer be able to rely on the availability of student-level data to conduct rigorous research. As such, alternatives must be sought in order to provide the research community with timely information about what works in education. Given the growing availability of publicly available school-level test score information, use of these data sources in lieu of student-level data is worth careful consideration.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

The overarching ELT evaluation seeks to estimate the difference between observed outcomes of ELT schools and the estimated outcomes of those schools if they had not received ELT on student tests scores in English Language Arts (ELA) and math. The more rigorous design options of a randomized control trial or a regression discontinuity design were not possible due to programmatic constraints. Therefore, the study team decided to employ a comparative short interrupted time series design (CISTS), which is a widely used quasi-experimental approach (Shadish, Cook, and Campbell, 2002). In this design, the introduction of the ELT program is modeled as an "interruption" in what would otherwise be assumed to be somewhat stable levels of test scores, and any changes observed in the scores of ELT schools before and after program implementation are compared to those changes observed in the matched comparison schools. The CISTS framework represents a rigorous quasi-experimental design that accounts for many alternative hypotheses that may confound the effect of ELT (although there might still be time-varying, school-specific, unmeasured variables related to study outcomes that could introduce bias into the estimated effect). See **Data Collection and Analysis** section for additional info on model.

In addition, moving to public-use school-level data necessitated careful consideration of how this would affect planned analyses. First, moving from a dataset with 30,000 student-level observations, to one with 300 school-level observations, would result in an inevitable loss of power[1] and potentially larger standard errors, which could affect our ability to detect differences between ELT and comparison schools. Second, without student-level data to capture within-school variation, we were concerned about our ability to adjust for clustering within schools.

[1] This concern was addressed by running power calculations with the assumption that data would be at the school level. Because the numbers of schools in our evaluation increases each year as additional cohorts are added, we anticipate having sufficient power to run student or school-level models by the third year of the evaluation. This may not be the case, however, for all evaluations and thus should be carefully considered before proceeding with a shift to a macro unit of analysis.

Third, when working with individual-level data we had control over which students to include or exclude in the analysis, as appropriate. Working with publically available school-level means would necessitate using whatever inclusion and exclusion rules used by the state. Finally, we had to consider whether it was theoretically appropriate to shift our evaluation from assessing the impact of ELT on *students* to the impact of ELT on *schools*.

Usefulness / Applicability of Method:
Demonstration of the usefulness of the proposed methods using hypothetical or real data.

Given the growing availability of publicly available school-level test score information, use of these data sources may aid researchers' ability to produce timely results on educational trends and interventions when student-level data are not available. Our results indicate that in certain situations, there are no substantive differences between school and student-level results.

Research Design:
Description of the research design.

(May not be applicable for Methods submissions)

N/A- see **Statistical, Measurement, or Econometric Model** section.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.
(May not be applicable for Methods submissions)

The study team requested 6th, 7th, and 8th grade student-level ELA and math achievement data from school districts for each ELT and matched comparison schools involved in the evaluation. Where available, data were obtained for up to five years prior to the start of ELT and one post-ELT year. School-level data were obtained for 6th, 7th, and 8th grade ELA and math scores from public websites maintained by districts and states for up to five years prior to the start of the ELT initiative and one post-ELT year.

Multiple regression models were used to estimate the effect of ELT on student achievement in ELA and Math, separately. For each subject, one model estimated the pooled effect across all grade-levels in which ELT was implemented while another estimated grade-specific effects. To standardize assessments that varied by state, test scores were z-scored within state, subject and grade. This process also facilitated the interpretation of the resulting effect estimates as effect sizes. All models were run both with student and school-level data. The models' key features include: 1) Predictor variable that marks observations in the ELT schools during the 2010-11 school year. In the pooled model, this was one primary predictor and results were pooled across grades. In the grade-specific model, there were predictor variables for grades 6, 7, or 8, to measure grade-specific effects of ELT 2) School fixed effects that absorb time-invariant factors specific to each school. 3) Year-by-grade fixed effects to capture changes in test scores due to year-specific factors (i.e., interrupted school year due to bad winter) that affected particular grades within all schools in a similar way 4) School-level covariates, including race/ethnicity. 5) School-level cluster-robust standard errors (also known as Hubert-White or "sandwich" standard errors; White, 1984; Liang & Zeger, 1986). In the *student-level analyses* these accounted for the correlation of outcomes of students in a given school in a given year (see the "cluster problem" as discussed in Schochet, 2008). In both the *student-level* and *school-level* analyses, these accounted for the serial correlation problem, which is due to the correlation of outcomes in a

given school across years (Bertrand, Duflo, & Mullainathan, 2004; Kezdi, 2004; and Angrist & Pischke, 2009). 6) *School-level analyses* were weighted by the number of students enrolled in each grade.

Findings / Results:

Description of the main findings with specific details.
(May not be applicable for Methods submissions)

Exhibit 1 compares the results of models that used student-level data to those that used school-level data. Model specifications for the two datasets were identical, with the exception that school-level models were weighted by the number of students enrolled in school. For each subject, two models were run, one that pooled effects across grades, and another that estimated effects separately for each grade.

(please insert Exhibit 1here)

The table highlights the similarity of the results between the two models, and, importantly, that conclusions would be the same whether based on the student or school-level data (no significant effects of ELT on ELA or math test scores in the first year of implementation). Theoretically, if the same students were present in both datasets, the point estimates would be identical regardless of the unit of analysis. That they are not highlights that decisions states make about the inclusion and exclusion of data in their publically available datasets are likely to be different than decisions made by researchers. Despite this, the point estimates between the student and school-level models were generally similar.

In addition, although we were concerned that the standard errors would be artificially inflated using school-level data, the results indicate that they too were generally comparable to those computed using student-level models. It is possible that the intra-class correlations at the schools we are studying are high, and thus there is less additional information to be gained from student-level data (and adjusting the standard errors for clustering at the student level).

Note that there were larger differences between the student and school-level models in the grade-specific models than the pooled models. Although we anticipate that by the third and final year of the study, we will be sufficiently powered to run the school-level models, these first year analyses were underpowered and dividing the impact estimate by grade likely further exacerbated this issue. In particular, the largest differences between the models are observed among 8th graders, where there were the fewest number of students in the evaluation.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

The ability to understand what works in education requires rigorous research methods which often rely on student-level data which have become increasingly difficult to obtain. In this paper we explore the methodological alternative of using publicly available school-level data instead of student-level data to conduct an impact analysis of a national ELT initiative on middle school students' test scores. Results comparing identical students-level models to school-level models were highly similar and conclusions would be the same whether based on student or school-level data. Thus, if the shift from student to school-level data is carefully weighed and methodologically appropriate for a given study, there are substantial benefits to shifting data sources.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Angrist, J. D., & Pischke, J.S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, N.J.; Oxford: Princeton University Press.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119 (1), 249-275.

Kézdi, G. (2004). Robust standard error estimation in fixed-effects panel models. *Hungarian Statistical Review*, 82(9), 95-116.

Liang, K. and Zeger, S.L.. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*. 73(1), 13-22.

Rubin, D.B. (1980). Bias reduction using Mahalanobis metric matching. *Biometrics*, 36, 293-298.

Rubin. D.B. (1997). Estimating causal effects from large data sets using propensity score. *Annals of Internal Medicine*, 127, 757-763.

Schochet, P. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, 33(1), 62-87.

Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton-Mifflin.

White, H. (1984). *Asymptotic theory for econometricians*. Orlando: Academic Press.

Appendix B. Tables and Figures

Exhibit 1: Comparison of Student and School-Level Models

	Student-Level Data			School-Level Data		
	Estimate	Std. Error	Sig	Estimate	Std. Error	Sig
ELA						
Pooled Model						
Effect of ELT pooled across grades	0.0658	0.1428	0.6452	0.0695	0.1107	0.5303
Grade Specific Model						
Effect of ELT on 6th Grade	0.0282	0.1799	0.8757	0.0386	0.1218	0.7511
Effect of ELT on 7th Grade	0.1970	0.1169	0.0921	0.1780	0.2159	0.4098
Effect of ELT on 8th Grade	0.1519	0.0885	0.0862	0.1346	0.1307	0.3030
Mathematics						
Pooled Model						
Effect of ELT pooled across grades	0.0791	0.1026	0.4405	0.1036	0.1068	0.3319
Grade Specific Model						
Effect of ELT on 6th Grade	0.0281	0.1078	0.7947	0.0606	0.0954	0.5254
Effect of ELT on 7th Grade	0.2319	0.1582	0.1427	0.2917	0.2916	0.3171
Effect of ELT on 8th Grade	0.2640	0.1554	0.0892	0.2046	0.3021	0.4982