

Program Value-Added: A Feasible Method for Providing Evidence on the Effectiveness of Multiple Programs Implemented Simultaneously in Schools

Robert Shand¹ , Stephen M. Leach²,
Fiona M. Hollands³, Florence Chang²,
Yilin Pan⁴, Bo Yan², Dena Dossett²,
Samreen Nayyer-Qureshi⁵, Yixin Wang³,
and Laura Head³

American Journal of Evaluation
1-23

© The Author(s) 2022



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/10982140211071017

journals.sagepub.com/home/aje



Abstract

We assessed whether an adaptation of value-added analysis (VAA) can provide evidence on the relative effectiveness of interventions implemented in a large school district. We analyzed two datasets, one documenting interventions received by underperforming students, and one documenting interventions received by students in schools benefiting from discretionary funds to invest in specific programs. Results from the former dataset identified several interventions that appear to be more or less effective than the average intervention. Results from the second dataset were counterintuitive. We conclude that, under specific conditions, program VAA can provide evidence to help guide district decision-makers to identify outlier interventions and inform decisions about scaling up or disinvesting in such interventions, with the caveat that if those conditions are not met, the results could be misleading.

Keywords

data analysis, impact evaluation, PreK–12 education, quasi-experiment

Introduction

Over the past few decades, federal education policy has increasingly emphasized the need for the scientific evaluation of educational programs (ESSA, 2015; NCLB, 2001). This push is at least partially

¹ American University, Washington, DC, USA

² Jefferson County Public Schools, Louisville, KY, USA

³ Teachers College, Columbia University, New York, NY, USA

⁴ World Bank, Washington, DC, USA

⁵ Columbia University, New York, NY, USA

Corresponding Author:

Robert Shand, American University, 4400 Massachusetts Ave NW, Washington, DC 20016, USA.

Email: rshand@american.edu

motivated by the fact that costs of educating students have risen dramatically over the past 50 years while educational gains have not risen proportionally (Deming & Figlio, 2016). The Every Student Succeeds Act (ESSA) explicitly defines what counts as evidence of effectiveness for educational programs, emphasizing both methodological rigor and also relevance to the local context. However, given the countless interventions in which American students participate each day, it is impractical to expect each one can be evaluated using the most rigorous research designs such as randomized controlled trials (RCTs) or quasi-experiments. These methods are resource-intensive, difficult to implement, and yet may still be uninformative (Lortie-Forgues & Inglis, 2019). Further, they represent a somewhat narrow definition of rigor that is well suited to answering a specific set of questions, but have other limitations and may not provide the types of evidence educational decision-makers need to inform their everyday decisions (Donaldson, Christie, & Mark, 2014). Additionally, such studies can only count among the top two tiers of ESSA evidence if the study population and context match sufficiently with those of the local decision-maker (USED, 2016). This requirement also presumes that education decision-makers are equipped to make such assessments although past research has demonstrated that evaluating research evidence requires capacity and infrastructure beyond that available in many education agencies (Honig & Coburn, 2008). But school and district leaders often seek data on program delivery and on student outcomes to help guide annual budgeting and planning (Levenson, Baehr, Smith, & Sullivan, 2014). The challenge is to provide information that is timely, practitioner-friendly, and still adequately rigorous.

In this paper, we test the hypothesis that an adapted and simplified form of value-added analysis (VAA), henceforth “program VAA,” can be usefully applied to provide acceptably rigorous estimates with readily available data that are more feasible to produce than full-scale experimental or quasi-experimental program evaluations. Since the idea of measuring teacher performance based on student learning gains was introduced by Hanushek (1971), value-added methods have been widely applied in teacher evaluation and, to a lesser extent, school and program evaluation. They have been used in a number of settings for high-stakes accountability measures, lower-stakes formative assessment, and for research purposes. The essence of VAA is an attempt to statistically isolate the effect of a teacher, school, or program on student learning by controlling for other factors that affect achievement outcomes, especially prior achievement results (Corcoran, 2010).

Rigor versus Feasibility Trade-off

Recent methods such as “rapid cycle evaluation” (e.g., Mathematica Policy Research, 2016) have been developed to produce local evidence more swiftly than traditional evaluation methods, but these still require significant analytic capacity and data collection, while also representing trade-offs relative to a full-scale evaluation. Furthermore, with school-based education decision-makers asserting limited confidence in their own ability to critically evaluate research evidence (May et al., 2020), there is a need for more feasible and locally relevant ways to evaluate educational programs. Such methods are likely to trade some degree of causal rigor for relevance (Hollands & Escueta, 2019), but may still meet the third tier of ESSA evidence. Locally relevant research evidence is more likely to influence decision-makers: Penuel and Farrell (2017) assert that local evaluations of programs may produce findings that are more “... *timely, credible and central to district leaders’ needs...*” and therefore more likely to be acted upon.

A further limitation in current educational research is that, although 80% of school district funds are spent on personnel (McFarland et al., 2017), few research studies attempt to estimate the impact of people per se on student achievement, with the exception of general estimates of teacher effects. Even when they do (e.g., Dobbie, 2011; Jackson, Rockoff, & Staiger, 2014; Rockoff, Jacob, Kane, & Staiger, 2011), there is little specification of the pedagogical skills and practices needed to produce observed improvements in student outcomes. Respected repositories of research evidence in

education such as the What Works Clearinghouse (WWC) and Evidence for ESSA (E4E) primarily feature studies of prepackaged interventions such as curricula, professional development, or educational technologies. Studies of the impact of assistant principals, instructional coaches, or counselors are not featured, and would be hard to execute because of the inability to randomly assign adults to jobs. This leaves large gaps in the evidence base needed to inform school and district decision-making about how to allocate budgets amongst programs and types of personnel.

As a result of this evidence gap, schools and districts must often produce their own evidence on the instructional programs they operate. This task typically requires striking a balance between evidence that is methodologically rigorous enough to guide and justify programmatic decisions while also being feasible to produce. For example, consider a cost-effectiveness analysis (CEA) employing both the ingredients method for cost estimation (Levin & McEwan, 2001) and a randomized experimental design to capture program effects. It is unlikely that many local education agencies (LEAs, e.g., school districts and equivalent entities that go by various names in different states) have the expertise and resources to conduct such a CEA but, even where possible, problems arise. First, LEAs are frequently either unable or unwilling to conduct RCTs; instead, they rely on various forms of quasi-experimental designs (QEDs) such as matched samples (e.g., Drits-Esser, Bass, and Stark, 2014). It can be difficult to establish a counterfactual condition because students are purposefully selected for interventions, and promising interventions must be provided to all students equitably. Second, the time-intensive nature of CEAs places practical limits on the number that can be executed. Third, CEA requires an estimation of the economic value of all resources required to implement a program, many of which are not easily observed because they represent resources shared across multiple activities.

Even when such research is conducted on a district's existing programs, the value for decision-making may be limited. LEAs are rarely prepared to shut down entrenched programs and are more likely to seek information about how to improve programs or how to tinker with the overall package of services provided to students (Yan & Hollands, 2018). An important feasibility consideration is the timeliness of results. In particular, the misalignment between end-of-year outcome data availability and fiscal year budget cycles may require reliance on interim or proxy outcome variables. In addition to potentially reduced statistical power (i.e., sample size may be reduced if interim measures are not mandated for all students) and/or reliability of measures, such choices may negatively bias effectiveness results when sufficient implementation time is not given (e.g., Roschelle et al., 2014). Thus, the challenge is providing education leaders with useful and timely evidence that is suitably credible to support programmatic and personnel decisions.

Literature Review and Conceptual Framework

Adapting Value-Added Analysis to Program Evaluation

Researchers have noted significant advantages of value-added models over the prior focus on absolute performance that did not account for differences in learners' starting positions, or in out-of-school factors that facilitate or impede learning (Jackson et al., 2014). While high-stakes accountability applications have been controversial, one promising use of the method is to identify outlier teachers, schools, programs, policies, or practices. Investigating those that are performing particularly well or poorly can lead to valuable lessons from high-performing teachers or programs and additional support for low performers (Braun, 2005).

In addition to both high-stakes teacher evaluation and lower-stakes uses for formative assessment of teaching, researchers have examined other potential applications of teacher and school value-added metrics to policy and practice. Much of this research has been synthesized by Jackson et al. (2014), including cross-validation of value-added measures based on their association

with other measures of teacher effectiveness. Value-added measures have been used as outcomes to assess the impact of teacher characteristics on effectiveness and to evaluate programs, policies, and interventions aimed at increasing teacher effectiveness (Dobbie, 2011; Rockoff et al., 2011). VAA has been applied in a variety of contexts, including Australia, Chile, the Netherlands, and Portugal, to measure the effectiveness of schools, variability in effectiveness of schools, and the effects of particular policies and practices within schools (Coates, 2009; Ferrão & Couto, 2014; Hofman, Hofman, Gray, & Wendy Pan, 2015; Milla, Martín, & Van Bellegem, 2016; Page, Martin, Orellana, & González, 2017; Thomas, 2001).

There has been significantly less research on potential uses of VAA for program evaluation despite suggestions for such applications (e.g., Meyer, 1997; National Research Council, 2010). Most commonly, in a natural extension given its extensive application in teacher evaluation, VAA has been used to evaluate teacher preparation, professional development, and coaching programs (e.g., Darling-Hammond, Newton, & Wei, 2010; Harris & Sass, 2011). These studies have found some ability of VAA to detect meaningful differences in policies and programs but suggest that VAA alone is insufficient to guide decisions. Another study analyzed literacy reform interventions in the context of broader school reform in San Diego, simultaneously comparing several programs using deviations from expected growth trajectories (Betts, Zau, & King, 2005).

The value-added approach is akin to the difference-in-differences estimator that has been widely applied in the evaluation of single programs and policy changes in that it estimates deviations from expected changes over time conditioning on initial values (see Abadie and Matias, 2018 for a discussion of recent innovations in this and related econometric methods and applications to program evaluation). A limitation of the difference-in-differences approach is that extemporaneous changes occurring alongside the programmatic or policy interventions under study can be confounding factors (Yu, 2013). In a real policy and decision-making context when experimental or quasi-experimental methods are impractical or not relevant to the decision at hand, a value-added approach for program evaluation, which is, in essence, comparative and simultaneous difference-in-differences for many programs at once, helps to address this concern by controlling for alternative programs in which students could be participating. It also involves controlling for prior test scores and student characteristics. Whether and under what conditions program VAA produces credible results are the primary subjects of this paper, using case study results from a large, urban school district in Kentucky.

Methodological Issues with VAA

Given their prominence, particularly in high-stakes teacher evaluation, value-added methods have been subject to extensive methodological research and debate. Koedel, Mihaly, and Rockoff (2015) offer an extensive synthesis of the main issues. Here, we highlight key findings from the methodological literature on teacher VAA to assess the suitability of applying a similar analysis for program evaluation purposes and to determine the best model specification for mitigating concerns about VAA methods. The objective of VAA is not to formally model the cumulative achievement function, but rather to use prior achievement and other covariates as proxies for omitted variables that could explain both assignments to treatment (teachers in most prior literature, and interventions in the present case) and outcomes. In other words, the critical questions are whether prior achievement and other covariates adequately serve as proxies for underlying academic ability and pre-treatment inputs into the learning process. This would control for any unobserved selection mechanism to obtain unbiased estimates of the unique contribution of one part of the educational process—teachers, schools, or interventions—in a manner that is suitably precise and reliable to inform policy decisions.

Both experimental and nonexperimental studies have found limited evidence of bias in well-specified VAA models. One experimental study tested how well nonexperimental VAA predicts student learning when students are randomly assigned to teachers (Kane, McCaffrey, Miller, & Staiger, 2013). Although bias was substantively small and statistically insignificant, the experiment had a small sample size and bias was measured with significant noise. A larger, quasi-experimental study assessed how actual changes in student learning compare with expected changes in value-added due to shifts in grade level taught by teachers (Chetty, Friedman, & Rockoff, 2014b). While the authors found little bias on average, they could not rule out errors for individual teachers. More recent quasi-experimental evidence in Ecuador, relying upon an alternating alphabetical assignment mechanism that is as good as random, found statistically and substantively significant differences in student performance by teacher, associated with teacher behaviors and parental perceptions of teacher quality (Araujo, Carneiro, Cruz-Aguayo, & Schady, 2016). Other research has suggested that bias is still a concern, finding evidence of “effects” on earlier test scores of subsequent teachers as evidence of systematic sorting, arguing that random assignment findings are too imprecise, are based on small samples, and are not generalizable. Critics also pose evidence that the grade switching used in the quasi-experimental approach is associated with student characteristics (Rothstein, 2010, 2017).

Finally, recent research has found that certain specifications of value-added models can find spurious “effects” of teachers on outcomes they could not possibly influence, such as student height (Bitler, Corcoran, Domina, & Penner, 2021). However, this seems to be driven by noise in the model, not bias, and can be alleviated with appropriate specification and the use of multiple pre-treatment years of data to increase precision. While there is some evidence to support the validity of teacher value-added measures, it is less clear how issues of bias and sorting apply to program evaluation when the assignment mechanism is even more likely to be purposeful rather than random; we examine these questions in this paper.

In addition to bias, there are also questions about the precision and reliability of value-added estimates. For instance, Minaya and Agasisti (2019) found that value-added measures of primary schools in Italy are robust to model specification but are not stable over time or across different outcome measures. Furthermore, they are noisy in the middle of the school performance distribution but more precise at the top and bottom, suggesting the model is better suited for identifying outlier performance and learning from it in a formative fashion than in strict ranking or other higher-stakes applications. This point—that the statistical properties of a value-added measure should be assessed relative to the stakes of a decision being made and compared with alternative sources of evidence in the absence of value-added—is at the heart of a discussion between the American Statistical Association (ASA) and several researchers. There is a general consensus that researchers and decision-makers should exercise some caution in using value-added measures, with an emphasis on improving schools, in conjunction with other measures, and being explicit about precision, noise, and assumptions while testing for sensitivity to these assumptions whenever possible (Chetty, Friedman, & Rockoff, 2014a; Morganstein & Wasserstein, 2014). There is disagreement in the degree of caution warranted: the ASA recommends avoiding high-stakes accountability applications of VAA due to the risk of bias and their correlational nature, while Chetty et al., note that ample experimental and quasi-experimental evidence suggests minimal bias and that a causal interpretation is warranted from a well-specified value-added model.

Methodological literature also provides guidance on the appropriate specification of a program value-added model to minimize these issues, though in many cases teacher VAA results are robust to a number of these decisions (Koedel et al., 2015). Key decisions include whether to perform a one-stage model including the desired effects (schools, teachers, or, in our case, interventions) as fixed or random effects or a two-stage model regressing the residuals from a growth model on a set of school, teacher, or intervention indicators; whether to specify the value-added estimates

themselves as fixed or random effects estimates; and whether to estimate value-added as the difference between two outcomes or if the pre-treatment outcome should be included as a covariate with the posttreatment outcome as the dependent variable. The preferred model is a one-stage model with a posttest on pretest rather than gain score approach, with little practical difference in whether the value-added effects themselves are specified as fixed or random (Koedel et al., 2015; Tekwe et al., 2004). Other important issues are how indicators are coded, selection of appropriate covariates to include in the model such as school and student fixed effects and multiple years of pre-treatment outcome data, Bayesian shrinkage, and procedures for addressing measurement error in the testing instrument (Aaronson, Barrow, & Sander, 2007; Ehlert, Koedel, Parson, & Podgursky, 2014; Herrmann, Walsh, & Isenberg, 2016; Kane et al., 2013).

Research Questions

In an attempt to assess whether and how program VAA can be used to conduct program evaluations that provide sufficiently rigorous, locally relevant, feasible, and timely evidence to inform school district decision-making, we set out to answer the following research questions:

- How can an adapted form of VAA be applied to analyze the comparative effect of specific programs, policies, and interventions on students' academic progress and other related outcomes?
- How well are the results of program VAA corroborated by extant sources of evidence?
- What conditions, in terms of student and program characteristics, selection or assignment mechanisms, measures, and availability of data, lend themselves to program VAA?

Setting and Context: Jefferson County Public Schools, KY (JCPS)

JCPS is a large, urban district located in Louisville, KY. With 94,466 students and 6,188 teachers in 168 schools, JCPS is the largest district in the state and one of the largest public school districts in the United States. In 2018–2019, JCPS's budget exceeded \$1.7 billion. The district's school-based decision-making (SBDM) model grants broad authority to each school's SBDM council with regard to hiring, budgeting, and policy-making (JCPS, 2018). SBDM councils typically consist of a school administrator and a small number of elected teachers and parents.

Over half of JCPS students come from minority populations (36% Black, 12% Hispanic/Latinx, and 9% Other). In the 2018–2019 school year, 60% of JCPS's schools received Title I funding, while more than 65% of students were eligible for free or reduced lunch and 5% were homeless. Nearly 10% of students were classified as English Learners and 13% required individualized education plans. Teachers were predominantly white (84%) and female (74%), with approximately 85% of all teachers holding a master's degree or higher and nearly 7% holding national board certifications. JCPS received 2 out of 5 stars in the Kentucky Department of Education's (KDE) accountability ratings (KDE, 2019). Based on state testing, reading proficiency at the elementary, middle, and high school levels was 45%, 50%, and 37%, respectively, while math proficiency was 40%, 35%, and 30.5% at those same levels. The average high school graduation rate was 83% and the dropout rate was 3%.

Data and Methods

We test the hypothesis that program VAA can be used to simultaneously evaluate a wide range of educational interventions using two independent sets of JCPS student-level data with information

about students' assignment to various interventions: Intervention Tab data and Investment Tracking System End-of-Cycle data, each described below. The two datasets feature different subgroups of students, types of interventions, and assignment mechanisms to interventions, and thus serve as two separate case studies for determining the conditions under which the method can be suitably applied to program evaluation.

For each case study, we apply a form of teacher VAA adapted to program evaluation purposes. We follow suggestions for model specification guided by the VAA literature and test for robustness whenever possible. In some cases, either the preferred specification in current research literature or a robustness check is not possible due to limitations in available data. Our method also shares similarities with a simultaneous difference-in-differences analysis of multiple treatments. However, we cannot test for parallel trends in pre-treatment outcomes because JCPS only began implementing the relevant outcome assessment at the start of the intervention year. As discussed in more detail below, we address concerns about potential selection effects in three ways: (1) by clarifying that the intent of this adapted method is to provide preliminary evidence on programs to suggest priority areas for further analysis; that is, it improves upon the status quo of limited information but should not be the sole influence on a program decision; (2) by taking advantage of features of a policy context and selection into treatment mechanism that minimizes this concern; and (3) by cross-validating our findings with prior research.

Data: Intervention Tab 2017–2018

One source of data on student-level participation in reading, math and behavior interventions at JCPS is a specialized dataset we refer to as the “Intervention Tab” which includes more than 70,000 student records per year. Starting in the 2014–2015 school year, KDE mandated that schools track all academic or behavioral interventions supported by KDE-specified funding mechanisms in the Intervention Tab of its Infinite Campus statewide student information system (KDE, 2020). In Academic Year 2017–2018, intervention services for each of the following four groups of students were mandated by the State of Kentucky to be recorded at the individual student level: K–3 students identified as requiring intervention services, students served by Extended School Services, students served with Mathematics Achievement Fund or Read to Achieve grants, and, for schools that have been classified as “Focus” schools for three or more years, all students who scored “Novice” (the lowest level) in state assessments.

In general, students receiving such interventions are more disadvantaged and have lower pre-treatment test scores than the district as a whole; the characteristics of students represented in the Intervention Tab data are described below. Schools are required to update student records at least four times per year. Separate student records are entered for regular year and summer school interventions with one entry for each combination of funding source, content area (e.g., reading, math, behavior), and tier (e.g., individual, small group). As a result, many students have multiple records. In addition, each record contains information on intervention duration, total hours, resolution, delivery method, and instructor type.

There are 32,207 students in the 2017–2018 Intervention Tab dataset with a large number of students receiving teacher-created interventions or interventions meant to address areas other than reading and math. For the purposes of this analysis, we focus on students who were involved in named, vendor-created interventions for reading or math, although we include teacher-created reading and math interventions as possible confounders. We do not include behavioral interventions as there are relatively few such records in the dataset and the data entry requirements set by the state only applied to a small set of behavioral interventions, increasing the risk of confounding effects with unobserved alternative treatments. We do not report results for teacher-created interventions as we do not have detailed information on what those interventions entail and thus any effects of these

interventions would not be distinguishable from the effectiveness of the teachers who created and implemented them and would not yield meaningful results for policy decisions about JCPS programs.

There were 8,886 unique students participating in vendor-created reading interventions and 5,677 unique students participating in vendor-created math interventions. The range of students participating in individual vendor-provided reading interventions was 4–2,757, with an average of 291 students. The range of students participating in vendor-provided math interventions was 1–1,225, with an average of 229 students. On average, a student in our restricted sample of vendor-created reading and math interventions was enrolled in 1.12 reading interventions and 1.06 math interventions. Table 1 summarizes descriptive statistics for students in the sample from the Intervention Tab dataset. On average, the sample includes somewhat more Black students and students with disabilities than the district at large. As anticipated, students in the Intervention Tab dataset had lower average, and more uniform, pre-treatment test scores than students in the district at large (see Appendix Figure A1 for detailed distributions).

Data: Investment Tracking System End-of-Cycle Items

We also tested whether program VAA methods can be useful to evaluate program and policy interventions supported by discretionary funds under the district’s Cycle-Based Budgeting System (Yan, 2017). School and district office leaders can submit budget requests for discretionary funds to supplement site-based budget allocations and those mandated by contract or under state and federal laws. These “investments,” most often to support personnel positions such as Success Coaches or Interventionists, are awarded on 1–5-year investment cycles and are tracked in JCPS’s Investment Tracking System. This system holds data on 20,928 students in 29 schools that received discretionary funds to support 47 investments that came up for budgetary review in 2018–2019. The student sample comprised elementary (49%), middle (10%), and high school (41%) grades, and was 46% female, 43% white, 35% Black, 12% Hispanic/Latinx, and 10% other race/ethnicity. The average number of students served by each investment item was 661 and the average number of investments per student was 1.48. Interventions targeted all students in a school; specific subgroups within a school, such as students with disabilities or students in a particular grade level and/or subject; or a specific target group of students based on identified need. In cases in which it was not clear which students were targeted by an intervention, we contacted school personnel for clarification and requested rosters to determine participation. Because results could be confounded by ongoing investments in comparison schools that were not up for budget review, and by similar interventions in other

Table 1. Descriptive Statistics for Intervention Tab Analytical Dataset.

Characteristic	Intervention Type	
	Reading	Math
No. of students	8,886	5,677
No. of interventions	37	28
Average number of students per intervention	290.5	228.5
Average number of interventions per student	1.12	1.06
Percent of students who are Black	51	50
Percent of students who are Hispanic/Latinx	12	11
Percent of students with disabilities	17	18
Average MAP Score 2017	174.51	183.18
Average MAP Score 2018	183.07	192.19

Note. MAP = Measures of Academic Progress.

schools that were funded by alternative sources and thus not appearing in the Investment Tracking System, we restricted our analysis in this dataset to schools that had active investments under the Cycle-Based Budgeting system and we included all current investment items as potential controls or confounders, though our analysis is focused on those investments up for end-of-cycle budget review.

Other Data: Administrative Data and Outcome Measures

We supplemented the student-level data on participation in interventions with administrative data on student characteristics—gender, race, grade level, disability status, and median income by Census tract—and three outcomes: fall 2017 and fall 2018 NWEA Measures of Academic Progress (MAP) scores, attendance rates, and days of suspension. We standardized MAP scores within grade levels for ease of interpretation and to minimize mechanical effects of the scaling of the exam.

Model Specification

We ran a series of models to test robustness to specification, but, in general, we followed guidance from Koedel et al. (2015) in using intervention fixed effects in a one-stage framework. In our preferred model, we regressed the 2018 outcome on the 2017 outcome, student characteristics, and a series of indicator variables for each intervention, with separate models for math and reading, as shown in Equation 1, where i indexes students in school s at time t , X represents the vector of student covariates, I is a vector of intervention indicator variables, and θ is a vector of the coefficients of interest indicating the effect of a given intervention conditional on the effects of other interventions, prior test scores, and student characteristics.

$$y_{ist} = \beta_0 + \beta_1 y_{ist-1} + X_{ist} \beta_2 + I_{ist} \theta + \varepsilon_{ist} \quad (1)$$

We standardized MAP scores within grade level and ran separate models for early elementary (K–2), upper elementary (3–5), and middle school interventions (6–8). We did not include high school interventions, as MAP administration was optional in JCPSS high schools and there is a high degree of non-ignorable missingness in both pretest and posttest outcomes. Running models separately by grade band, in addition to standardizing test scores within grade level, mitigates concerns about mechanical effects due to the scaling of the assessment, lower average growth at higher grade levels, and the fact that even though we are using MAP as an outcome for all interventions, different interventions are likely to target different skills and levels of student ability.

We ran a series of robustness checks with alternative specifications: pooled across all students versus by grade band, with and without student characteristics as covariates, with and without school fixed effects, and with growth as a dependent variable. Because we are testing several hypotheses about programs at once, we considered the possibility of skewed results due to multiple inferences. We report conventional *SE* and hypothesis tests in our main results, as the objective of this analysis is to determine potential outliers for further study. Thus, substantively meaningful effects that are measured less precisely and have higher probabilities of a Type I error are still of potential interest. However, to test the robustness of our results to a more stringent criterion we also applied a Benjamini–Hochberg correction, reported in Online Appendix Table A1 (Benjamini & Hochberg, 1995). We also report robustness checks using an alternative, two-stage model specification in Online Appendix C.

Application to Intervention Tab Data

For the Intervention Tab dataset, we minimized selection bias by restricting our sample to the students who received a vendor-created reading or math intervention. Students in the Intervention Tab dataset are more similar to other students in the dataset than those outside the dataset, implying that selection into any intervention is a more serious issue than which particular intervention a student received and thus the method of restricting the analysis to students receiving any intervention does at least partially mitigate selection bias. Additionally, due to mechanical reasons (e.g., time constraints), if students are receiving an officially logged intervention, they are less likely to be receiving other unobserved interventions that would potentially bias effects of observed interventions downward. We effect coded interventions so their effects could be interpreted relative to the overall average effect of all interventions in the dataset (Mihaly, McCaffrey, Lockwood, & Sass, 2010), which is marginally lower than but almost equivalent to nationally normed growth on the MAP assessment (NWEA, 2020). Effect coding has a sum-to-zero constraint which means that students can only be in one “intervention.” We, therefore, estimated effects for all individual interventions and unique combinations of interventions in which any given student participated.

Application to Investment Tracking System Data

Due to the different nature of the dataset and the types of interventions, students, and schools included, our approach to the analysis of Investment Tracking System End-of-Cycle items was somewhat different. We did restrict the sample to schools that had any investments in the Investment Tracking System to avoid comparisons with schools that may have had similar investments but from alternative funding sources which would not have been captured in the Investment Tracking System. We included all students in all schools which had at least one active investment item in the Investment Tracking System, regardless of whether the students were the intended beneficiaries of the investment(s). This provided a suitable comparison group of students who are not identified as targets of any particular investments. We used dummy coding so that all coefficients are interpreted as relative to students receiving no intervention. From a policymaking perspective, the question of whether an investment strategy works across multiple sites is of greater interest than, for instance, whether funding an assistant principal at one particular school improves student achievement, so we combined similar or identical investments across schools as a single investment and included school fixed effects in our preferred models to capture school-level heterogeneity.

Cross-Validation of Results

As a means of assessing the validity of the Intervention Tab results, we searched for existing studies on the math and reading items for which significant positive or negative results were obtained. We initially searched for studies in the WWC and E4E databases of educational interventions. We considered these the most reliable sources of corroborating evidence as they establish high standards of methodological rigor for accepting studies as evidence on the effectiveness of educational interventions (e.g., WWC, 2020). If we found studies in E4E or WWC, we report these results and do not report results of any additional studies. If we did not find an intervention listed in either of these two databases, we searched other sources including Education Endowment Fund, ERIC, a university library, and Google Scholar. Results from these alternative sources of evidence are presented in Online Appendix Table E1. If we found only one study that we deemed reliable (based on an RCT, QED, or correlational study conducted by an independent evaluator), we report information from this one study. If multiple studies with credible designs were found, we provide a summary of the findings. For interventions with no high-quality evaluations, we looked for studies on the

vendor's website and conducted an open Internet search which could surface other reports or evaluations, and summarize the available evidence.

Findings

Intervention Tab

Our findings indicate that relatively few interventions have statistically significant value-added that differ substantially from average growth. However, we do find notable outliers worthy of further investigation. Due to the sheer number of interventions and combinations of interventions, for parsimony, we only report here results for 15 math and 8 reading interventions that are statistically significant in at least one of our models. Full results including those that were not statistically significant, as well as the results with the Benjamini–Hochberg correction, are available in Online Appendix Table A1. Table 2 summarizes the results of our models by subject area and grade band, with coefficients representing standardized deviations from average growth among all students receiving vendor-provided reading and math interventions on respective MAP scores. Our preferred model regressed post-intervention MAP scores on pre-intervention scores and student covariates, and Models 2, 3, and 4 excluded student covariates and used growth as a dependent variable with and without student covariates, respectively.

Among K–2 elementary reading programs, we see positive and significant effects in our preferred model at the 0.05 level or below of Lexia Reading, Education Galaxy Reading, and the Edmentum Reading Suite, and marginally significant effects at the 0.1 level of the combination of Lexia and Reading Recovery, Leveled Literacy Intervention, and the combination of comprehensive intervention model (CIM) Reading Recovery and Lexia Reading. Results are of very similar magnitudes and patterns of statistical significance across model specifications. There are fewer statistically significant interventions in reading in upper elementary and middle school grades in our preferred model, with iLit and MobyMax Reading both having statistically significant below-average effects relative to all other interventions. For elementary school math, we find above-average and significant effects of the combination of Do the Math by Marilyn Burns with Think Through Math and Study Island Math; positive and marginally significant effects of the combination of Do the Math by Marilyn Burns with Envision Math Pearson; Engage New York Math, Reflex Math, and Teaching Student-Centered Mathematics by Van de Walle; and statistically significant below-average effects of Add+VantageMathRecovery (AVMR) with MobyMax Math; DreamBox with KCM Comp Intermediate I, and Go Math. We do not observe any statistically significant differences from average MAP performance among middle school math programs in our preferred model, but in alternative models with similar point estimates to our preferred model we see significant positive effects of JCPS Math eSchool, marginally significant positive effects of Edgenuity Math, and significant negative effects of Coach Books Math combined with Engage New York Math. The number of statistically significant estimates is well in excess of the 5% we would expect by chance alone, although only seven interventions remain significant with the Benjamini–Hochberg correction as reported in Online Appendix Table A1.

End-of-Cycle Investments

For the Investment Tracking System End-of-Cycle investments, effects are interpreted relative to growth for students who are in schools that had investments in the system but who were not themselves targeted by any investments. We observe statistically significant negative effects on reading scores for school nurses, career readiness programs, shuttle drivers, summer literacy programs, assistant principals, interventionists, and mental health counselors, and statistically significant positive

Table 2. Summary of Intervention Tab Interventions with Statistically Significant Results in Program Value-Added (VAA) Model.

Intervention	Preferred Model	Model 2	Model 3	Model 4
Early Elementary (K-2) Reading				
Lexia Reading	0.26** (0.087)	0.23** (0.087)	0.20* (0.10)	0.24* (0.10)
Lexia + Reading Recovery	0.53~ (0.30)	0.69* (0.31)	0.67* (0.34)	0.42 (0.33)
Leveled Literacy Intervention	0.199~ (0.10)	0.23* (0.10)	.19 (0.11)	0.24* (0.011)
Education Galaxy Reading	0.39* (0.19)	.49* (0.19)	0.07 (0.21)	.12 (0.20)
Edmentum Reading Suite	0.25* (0.10)	.19~ (0.10)	.24* (0.11)	.24* (0.12)
CIM Reading Recovery + Lexia	0.75~ (0.43)	0.80~ (0.42)	0.49 (0.48)	0.35 (0.46)
Upper Elementary (3-5) Reading				
MobyMax Reading	-0.16* (0.07)	-0.16* (0.07)	-0.17* (0.07)	-0.07* (0.07)
Middle School (6-8) Reading				
iLit	-0.35** (0.12)	-0.35** (0.13)	-0.27* (0.13)	-0.24~ (0.14)
Early Elementary (K-2) Math				
AVMR + MobyMax Math	-0.67** (0.25)	-0.78** (0.25)	-0.64* (0.28)	-0.58* (0.27)
Do the Math by Marilyn Burns + Envision Math Pearson	0.39~ (0.24)	0.18 (0.23)	0.48~ (0.25)	0.54* (0.25)
Engage New York Math	0.34~ (0.02)	0.31 (0.20)	0.23 (0.22)	0.26 (0.22)
Upper Elementary (3-5) Math				
Do the Math by Marilyn Burns + Think Through Math	0.35* (0.17)	0.39* (0.17)	0.36* (0.17)	0.34* (0.18)
DreamBox + KCM Comp Intermediate I	-0.57* (0.25)	-0.53* (0.25)	-0.62* (0.26)	-0.59* (0.26)
Go Math	-0.42* (0.19)	-0.49* (0.19)	-0.45* (0.20)	-0.38* (0.20)

(continued)

Table 2. Continued.

Intervention	Preferred Model	Model 2	Model 3	Model 4
Reflex Math	0.1~ (0.06)	0.09 (0.06)	0.11~ (0.06)	0.13* (0.06)
Study Island Math	0.14* (0.06)	0.09 (0.06)	0.14* (0.06)	0.16** (0.06)
Teaching Student-Centered Mathematics by Van de Walle	0.22~ (0.12)	0.20~ (0.12)	0.27* (0.12)	0.24~ (0.13)
Middle School (6–8) Math				
Coach Books Math + Engage New York Math	–0.44 (0.18)	–0.43* (0.19)	–0.48* (0.20)	–0.36~ (0.20)
JCPS Math eSchool	0.19 (0.12)	0.23* (0.12)	0.23~ (0.13)	0.26* (0.13)
Edgenuity Math	0.10 (0.08)	0.12 (0.08)	0.11 (0.08)	0.15~ (0.08)
Posttest on pretest	X	X	X	X
Growth as Dependent Variable (posttest–pretest)	X	X	X	X
Student covariates	X	X	X	X

Note. SE is shown in parentheses.]CPS = Jefferson County Public Schools.
 ~p < 0.1; *p < 0.05; **p < 0.01.

Table 3. Summary of Alternative Outcome Reanalysis of Preferred Models (1-Year vs. WMA Pretest).

Grades	Subject	Statistically significant interventions in model		Statistically significant student characteristics		R^2 1-year	R^2 WMA
		1-year	WMA	1-year	WMA		
<i>Suspensions</i>							
K–2	Reading	2	2	2	2	.22	.22
3–5		3	2	2	1	.22	.25
6–8		1	3 ^a	2	2	.39	.38
K–2	Math	1	1	3	3	.19	.19
3–5		2	2	1	1	.23	.26
6–8		1	1	1	1	.37	.38
<i>Attendance</i>							
K–2	Reading	1	2 ^a	5	6	.41	.39
3–5		5	3	1	0	.45	.47
6–8		1	1	0	0	.60	.61
K–2	Math	2	2	1	1	.59	.52
3–5		1	1	2	2	.48	.53
6–8		0	2 ^a	1	1	.56	.54

Note. WMA = weighted moving average; ^aSignificant interventions in WMA model are not a subset of those in 1-year model.

effects for school resource officers. We find very similar patterns of effects for math scores and for suspensions (with a positive sign indicating a higher number of occurrences or days of suspensions and thus being interpreted as a negative effect). For reasons discussed in more detail below, and suggested by the strongly counterintuitive nature of these findings, we believe these results are subject to a number of biases and thus serve as a useful guide to circumstances under which this method will not yield useful results for policy. We present the full results in Online Appendix Table B1 but do not include them here to caution against overinterpreting these findings.

Robustness Checks: Alternative Outcomes and Multiple Years of Pre-Treatment Data

In an effort to test the robustness of our preferred models, we reran Intervention Tab models using data from multiple years on two alternative student-level outcomes: attendance and out-of-school suspension occurrences (as previously noted, we were unable to use multiple years of MAP data as this test was only introduced at JCPS in 2017–2018). Although the interventions included in our study were primarily focused on improving students' math or reading skills, there is some evidence that such interventions may lead to increased student engagement (e.g., Kim et al., 2020), which is, in turn, associated with improved behavioral outcomes (e.g., Quin, Heerde, and Toumbourou, 2018). Substituting attendance and suspensions as alternative outcomes permitted us to compare results for each outcome using a single year of pretest values versus including data from multiple prior years. Specifically, we computed a 3-year weighted moving average (WMA) for each outcome using weights of .20, .30, and .50 for Years 1, 2, and 3, respectively. Because we did not expect early elementary or mobile students to have 3 years' worth of prior data, in an effort to maximize the sample size, weights of (.30, .70) and (1.00) were used to calculate WMA for students with 2 years and 1 year, respectively, of prior outcome data.

Table 3 provides a summary of the model results for each combination of alternative outcome and pretest approach. With two exceptions, the regression coefficients on the WMA pretest values were

Table 4. Summary of Alternative Evidence from WWC and E4E.

Combinations of Intervention	Preferred Model	Individual Intervention	Alternative Evidence Findings: WWC		Alternative Evidence Findings: E4E	
			Domain Area	Average Effect Size	ESSA Rating	Average Effect Size
Lexia Reading	0.26	Lexia	Alphabetics Fluency Comprehension General Reading Achievement	+ 0.27 na + 0.22 ns + 0.27 ns + 0.23 ns	Promising	+ 0.28
Lexia + Reading Recovery (RR)	0.53	Lexia RR	See above Positive effect with no overriding contrary evidence		Strong	+ 0.43
Leveled Literacy Intervention (LLI)	0.199	LLI	Alphabetics Fluency Comprehension General Reading Achievement	+ 0.55 na + 1.71 sig. + 0.36 na + 0.75 na	Strong	+ 0.13
Education Galaxy Reading Edmentum Reading Suite	0.39 0.25	Education Galaxy Edmentum Reading Suite	Positive effects Alphabetics Fluency General Reading Achievement	+ 0.13 na + 0.27 na + 0.27 na	No studies met inclusion requirements Not found	
CIM Reading Recovery + Lexia	0.75	CIM	Not found		No studies met inclusion requirements	
MobyMax Reading	-0.16	RR Lexia MobyMax Reading	See above See above Not found		No studies met inclusion requirements	
iLit	-0.35	iLit	Not found		Qualifying studies found no significant positive outcomes.	

(continued)

Table 4. Continued.

Combinations of Intervention	Preferred Model	Individual Intervention	Alternative Evidence Findings: WWC		Alternative Evidence Findings: E4E	
			Domain Area	Average Effect Size	ESSA Rating	Average Effect Size
AVMR + MobyMax Math	-0.67	AVMR MobyMax Math	Not found Not found		Not found No studies met inclusion requirements	
Do the Math (DtM) by Marilyn Burns + Envision Math Pearson	0.39	DtM	Not found		No studies met inclusion requirements	
Engage New York Math	0.34	Envision Math	No studies meet WWC group design standards		Qualifying studies found no significant positive outcomes	
Do the Math (DtM) by Marilyn Burns + Think Through Math (TTM)	0.35	Engage NY DtM	Not found See above		Not found	
DreamBox + KCM Comp Intermediate I	-0.57	TTM DreamBox	Not found Potentially positive effects Mathematics Achievement	+ 0.11 sig.	Not found Strong	+ 0.11
Go Math Intervention Materials	-0.42	KCM Go Math	Not found Not found		Not found Qualifying studies found no significant positive outcomes	
Reflex Math	0.1	Reflex Math	Not found		No studies met inclusion requirements	
Study Island Math	0.14	Study Island	Not found		No studies met inclusion requirements	
Teaching Student-Centered Mathematics by Van de Walle	0.22	Van de Walle	Not found		No studies met inclusion requirements	
Coach Books Math + Engage New York Math	-0.44	Coach Books Math	Not found		Not found	
JCPS Math eSchool	0.19	Engage NY JCPS Math eSchool	See above Not found		Not found	
Edgenuity Math	0.10	Edgenuity	Potentially positive effects Mathematics Achievement	+ 0.31 na	Qualifying studies found no significant positive outcomes	

Note. E4E = Evidence for ESSA; ESSA = Every Student Success Act; JCPS = Jefferson County Public Schools; WWC = What Works Clearinghouse.

larger in magnitude than those of the single-year pretest models (Δ suspension = .10; Δ attendance = .22). With four exceptions (see rows with subscript “a” in Table 3), significant interventions in the WMA models were a subset of those in the 1-year models with similar or identical coefficients. In three of those exceptional cases, the WMA models produced one or two additional statistically significant interventions compared to the respective 1-year models. In the fourth case, middle school math with attendance as the outcome, the set of three significant interventions in the WMA model were not a subset of the five significant interventions in the 1-year model. On the other hand, WMA models tended to produce fewer statistically significant student characteristics than corresponding single-year pretest models, with coefficients generally equal or smaller in magnitude but in the same direction in all cases. As expected due to limited previous outcome data, early elementary (grades K–2) models were more similar in number and magnitude of statistically significant regression coefficients than models for later grades that had more years of prior pretest data available. Detailed results with actual coefficients on each intervention are available in Online Appendix Table D1.

Comparison of Findings on Intervention Tab Items to Extant Research

Table 4 and Online Appendix Table E1 summarize the results of our search for alternative research findings with which to corroborate our own findings. Table 4 only shows results from WWC and E4E while the Online Appendix includes additional sources of evidence. For a few of the reading interventions—Lexia Reading, Reading Recovery, and Leveled Literacy Intervention—existing studies that qualified for WWC or E4E corroborated our positive findings quite well both in direction and magnitude. For Education Galaxy Reading and Edmentum Reading Suite, other sources of evidence supported our positive findings. Existing evidence on MobyMax Reading was mixed while our analysis found a small negative effect. For iLit, we also found a negative effect while the vendor refers to several studies that show positive effects. We could not locate these studies. Overall, we found no rigorous, third-party studies that contradicted our findings for reading interventions.

Corroborating our findings on the math interventions with existing evidence was harder, partly because so few of the interventions were listed in WWC or E4E and several had no evidence of any kind on their effectiveness, and partly because several were delivered in combination with another math intervention. While only two of the eight reading interventions we report on were combinations of two interventions, five of the 12 math interventions were combinations. In these instances, we could not determine which individual intervention might be contributing to the effect and how much. We could only find high-quality evidence in WWC or E4E for one of the interventions in one of the math combinations (DreamBox), and for one of several programs that constitute another math intervention, Edgenuity Math. In the latter case, the WWC study corroborated our finding in direction (positive) but the WWC effect was higher in magnitude. In the case of the combination, the WWC/E4E effects were positive for one of the interventions and our effect for the combination was negative.

For all other math interventions, we needed to look beyond WWC and E4E for alternative sources of evidence. For MobyMax Math, a dissertation and a vendor study found positive effects while we found a negative effect for the combination of MobyMax Math and AVMR. We found no alternative evidence on AVMR. Our positive effects for the combinations of Do the Math by Marilyn Burns with Envision Math Pearson or with Think Through Math were corroborated by alternative sources. For Go Math Intervention Materials we found a moderate negative effect but no studies on the intervention materials per se to corroborate this. Studies on Go Math! as a whole found negligible effects. Our small positive finding for Reflex Math was generally supported by a variety of dissertations and vendor studies. Alternative sources on Study Island showed mixed results compared with our small positive finding. We found no alternative evidence at all on Engage New York Math,

Teaching Student-Centered Mathematics by Van de Walle, or JCPS Math eSchool to corroborate our positive findings. We found no alternative source of evidence on Coach Books Math to corroborate our negative finding for its combination with Engage New York Math, or of KCM Comp Intermediate I to corroborate our negative finding for its combination with DreamBox.

Discussion

Summary of Results

We conclude that, in some situations, program value-added can be a useful complement to other sources of evidence to inform decisions about whether to continue, scale, or replicate programs. In particular, there is a promise for identifying outlier interventions and investments that merit further investigation. The method suffers from some limitations that imply that it should not be the sole input for high-stakes decisions and that policymakers should assess whether it is worthwhile to gather the necessary data. Because it is not possible to observe the counterfactual of all possible “interventions” in which students might participate and because students are not randomly sorted into interventions, there is likely selection bias, most frequently in the downward direction. Further, as with teacher value-added estimates, even with many observations, there are power and reliability concerns due to measurement error, collinearity due to similar patterns of interventions, small numbers of students participating in individual interventions, and a large number of parameters estimated relative to the sample size.

Overall, the alternative, existing sources of evidence supported our Intervention Tab results for reading interventions but did not serve particularly well to either support or challenge our analysis of math interventions. Perhaps the bigger concern is that so little evidence can be found on so many of the math interventions being used in a large school district.

Differences in Intervention Tab and End-of-Cycle Results

Our findings suggest that our adaptation of value-added methods to program evaluation may be more promising under some conditions than others. Our Intervention Tab analysis of reading interventions was generally corroborated by extant evidence, but the end-of-cycle analysis yielded findings that were highly counterintuitive and contrary to expectations based on theory and prior research. We contend that these differences in findings suggest more general conditions for context, program assignment, and data under which adapting VAA to program evaluation is more valid and useful. Analysts and decision-makers should exercise caution in presenting and interpreting findings if these conditions are not met, as results could be misleading due to selection bias or measurement error.

First, our analysis of Intervention Tab programs was restricted to a well-defined set of students all of whom were receiving state-mandated interventions; therefore, the comparison group for students receiving any one intervention (or combination) was students receiving other interventions in the dataset. This is still an imperfect solution to providing a counterfactual because, even within this restricted sample, there are a wide range of students with different academic strengths and challenges, and interventions targeting different skills and subskills. Furthermore, this analysis can only provide relative effects rather than absolute measures of program effectiveness. However, it helps to mitigate selection bias by comparing only students who were receiving interventions. Further, students whose interventions were tracked in the system were less likely to participate in other interventions we could not observe, which would bias results downward due to unobserved treatments received by comparison group students.

On the other hand, interventions that we assessed from the Investment Tracking System were likely more subject to selection bias in that the students served by an intervention were lower-performing than students in the comparison group, and perhaps less likely to show improvements in the measured outcomes. Many interventions in the Investment Tracking System were school-wide but, for others, it was not clear exactly which students were targeted. As a result, effects of school-wide programs could be confounded with the effects of the schools themselves or the types of schools seeking out such investments. Additionally, it was not always clear if investments were truly new programs or simply represented a shift in how those programs were paid for, for example, from the site-based budget to discretionary funding. Other schools may very well have had similar investments such as nurses or assistant principals but paid for them out of other sources rather than submitting budget requests via the Investment Tracking System. This suggests that particular features of the setting, intervention targeting, and assignment mechanism, along with features of the data, are amenable to this type of analysis. Interventions that target specific subgroups of students so that students being compared either between treatments or between a treatment and comparison group are more similar to one another at the outset and datasets with relatively exhaustive lists of interventions and students would be the most likely candidate for this analysis to yield credible and useful estimates, at least of relative program effectiveness. These estimates could then be used to target further examination of what outlier positive programs are doing well and whether negative outlier programs should be tweaked or eliminated.

Conditions for Usefulness of Evidence From Value-Added Analysis

The findings in the current study illustrate that the value-added modeling approach can provide credible evidence on the effectiveness of programs when a LEA (1) implements a large set of interventions and programs while tracking student-level participation in each, (2) administers a district-wide common assessment, such as MAP, and (3) uses program evaluations to inform decisions about how to invest budgets. Under these conditions, the value-added approach can, in a single analysis, provide evidence on the relative effectiveness of a large number of programs and interventions and focus attention on outlier programs that merit further investigation. These results may support larger-scale implementation of relatively effective programs, or consideration of alternative programs to replace relatively less effective programs.

Limitations and Further Research

Although this study provides promising preliminary evidence for conditions under which adapting VAA may be appropriate and useful for program evaluation purposes, further study is needed on model specification, how the model performs with multiple years of prior data, and ways to mitigate the effects of unobserved counterfactual programs, given that limitations in available data in this context did not allow for full exploration of each of these areas. While we have developed hypotheses for the usefulness of this approach and the conditions under which it is most applicable and feasible, these rely upon two case studies in one school district. Future work will more formally test the hypothesized optimal conditions for this method using simulated data and consider whether the considerable data collection and analysis work are worthwhile for a district. It will also be important to investigate qualitatively how such evidence enters into the decision-making process and whether shifting resources into programs that have scored higher value-added metrics results in better student performance.

Conclusion

In sum, since the ratification of ESSA, there is a renewed focus on ensuring that schools are investing in evidence-based programs and interventions. VAA provides an approach in which education agencies can examine multiple programs and interventions simultaneously, taking into account prior outcome measures and student characteristics to mitigate concerns about selection bias. Our analysis of two comprehensive datasets which document student-level participation in interventions provides early evidence for conditions under which this method is more and less promising. Such evidence arguably constitutes Tier III evidence for ESSA, being based on correlational methods and undeniably relevant to the local context. While it should not be the sole approach in program evaluation, it can provide evidence to identify outlier programs and interventions which should be further investigated.

Acknowledgments

We are grateful to Joseph Prather, retired from Jefferson County Public Schools, for invaluable assistance in assembling the data, and to Minetre Martin of American University for help with assembling the final manuscript. We also appreciate comments from Tommaso Agasisti, Mara Soncin, and participants in the 2020 Institute for Education Sciences Principal Investigators' conference and the Spring 2020 Association for Education Finance and Policy conference, where earlier versions of this work were presented.


Funding

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305H180003 to Teachers College, Columbia University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iD

Robert Shand  <https://orcid.org/0000-0001-8151-3913>

Supplemental Material

Supplemental material for this article is available online.

References

- Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, 25(1), 95–135.
- Abadie, A., & Matias, D. C. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, 10, 465–503.
- Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics*, 131(3), 1415–1453.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289–300.
- Betts, J. R., Zau, A., & King, K. (2005). *From blueprint to reality: San Diego's education reforms*. San Francisco, CA: Public Policy Institute of California.

- Bitler, M., Corcoran, S., Domina, T., & Penner, E. (2021). Teacher effects on student achievement and height: A cautionary tale. *Journal of Research on Educational Effectiveness*, 14(4), 900–924. <https://doi.org/10.1080/19345747.2021.1917025>
- Braun, H. I. (2005). Using student progress to evaluate teachers: A primer on value-added models. Policy Information Perspective. *Educational Testing Service*.
- Chetty, R., Friedman, J., & Rockoff, J. (2014a). Discussion of the American Statistical Association's statement (2014) on using value-added models for educational assessment. *Statistics and Public Policy*, 1(1), 111–113.
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9), 2593–2632.
- Coates, H. (2009). What's the difference?: A model for measuring the value added by higher education in Australia. *Higher Education Management Policy*, 21, 1–20.
- Corcoran, S. P. (2010). Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice. Education policy for action series. *Annenberg Institute for School Reform at Brown University (NJI)*.
- Darling-Hammond, L., Newton, X., & Wei, R. C. (2010). Evaluating teacher education outcomes: A study of the Stanford teacher education programme. *Journal of Education for Teaching*, 36(4), 369–388.
- Deming, D. J., & Figlio, D. (2016). Accountability in US education: Applying lessons from K–12 experience to higher education. *The Journal of Economic Perspectives*, 30(3), 33–55.
- Dobbie, W. (2011). Teacher characteristics and student achievement: Evidence from Teach for America. Unpublished manuscript, Harvard University.
- Donaldson, S. I., Christie, C. A., & Mark, M. M. (2014). *Credible and actionable evidence: The foundation for rigorous and influential evaluations*. Thousand Oaks, CA: Sage Publications.
- Drits-Esser, D., Bass, K. M., & Stark, L. A. (2014). Using small-scale randomized controlled trials to evaluate the efficacy of new curricular materials. *CBE—Life Sciences Education*, 13(4), 593–601.
- Ehlert, M., Koedel, C., Parson, E., & Podgursky, M. (2014). The sensitivity of value-added estimates to specification adjustments: Evidence from school-and teacher-level models in Missouri. *Statistics and Public Policy*, 1, 19–27.
- ESSA (2015). Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).
- Ferrão, M. E., & Couto, A. P. (2014). The use of a school value-added model for educational improvement: A case study from the Portuguese primary education system. *School Effectiveness and School Improvement*, 25, 174–190.
- Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, 61(2), 280–288.
- Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7-8), 798–812.
- Herrmann, M., Walsh, E., & Isenberg, E. (2016). Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels. *Statistics and Public Policy*, 3, 1–10.
- Hofman, R. H., Hofman, W. A., Gray, J. M., & Wendy Pan, H. (2015). Three conjectures about school effectiveness: An exploratory study. *Cogent Education*, 2(1), 1006977.
- Hollands, F. M., & Escueta, M. (2019). How research informs educational technology decision-making in higher education: The role of external research versus internal research. *Educational Technology Research & Development*, 68(1), 163–180. <https://doi.org/10.1007/s11423-019-09678-z>
- Honig, M. I., & Coburn, C. (2008). Evidence-based decision making in school district central offices: Toward a policy and research agenda. *Educational Policy*, 22(4), 578–608.
- Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics*, 6(1), 801–825.
- Jefferson County Public Schools (2018). *School-based decision-making*. <https://www.jefferson.kyschools.us/about/leadership/sbdrm>.

- Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Research Paper. MET Project. Bill & Melinda Gates Foundation.
- Kentucky Department of Education (2019). *Kentucky's school report card*. <https://www.kyschoolreportcard.com/home?year=2019>.
- Kentucky Department of Education (2020). *Infinite campus intervention tab*. https://education.ky.gov/educational/int/ksi/Pages/ksilC_InterventionTab.aspx.
- Kim, J. S., Burkhauser, M. A., Mesite, L. M., Asher, C. A., Relyea, J. E., Fitzgerald, J., & Elmore, J. (2021). Improving reading comprehension, science domain knowledge, and reading engagement through a first-grade content literacy intervention. *Journal of Educational Psychology, 113*(1), 3. <https://doi.org/10.1037/edu0000465>
- Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180–195.
- Kraft, M. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253.
- Levenson, N., Baehr, K., Smith, J. C., & Sullivan, C. (2014). *Spending money wisely: Getting the most from school district budgets*. District Management Council. http://smarterschoolspending.org/sites/default/files/resource/file/Research_Spending%20Money%20Wisely.pdf.
- Levin, H. M., & McEwan, P. J. (2001). *Cost-Effectiveness analysis: Methods and applications* (2nd ed). Thousand Oaks, CA: Sage Publications.
- Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher, 48*(3), 158–166.
- Mathematica Policy Research (2016). *Rapid-cycle evaluation*. <https://www.mathematica.org/our-publications-and-findings/publications/rapid-cycle-evaluation>.
- May, H., Farley-Ripple, E., Blackman, H., Wang, R., Tilley, K., & Shewchuk, S. (2020). *Individual and school-level capacity to critically evaluate research: A multilevel organizational analysis*. Poster session at the Annual Meeting of the Society for Research on Educational Effectiveness, Washington, DC.
- McFarland, J., Hussar, B., de Brey, C., Snyder, T., Wang, X., & Wilkinson-Flicker, S., ... S. Hinz (2017). *The condition of education 2017 (NCES 2017- 144)*. U.S. Department of education. Washington, DC: National Center for Education Statistics, <https://nces.ed.gov/pubsearch/pubinfo.asp?pubid=2017144>.
- Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review, 163*, 283–301.
- Mihaly, K., McCaffrey, D., Lockwood, J. R., & Sass, T. (2010). Centering and reference groups for estimates of fixed effects: Modifications to felsdvreg. *The Stata Journal, 10*, 82–103.
- Milla, J., Martín, E. S., & Van Belleghem, S. (2016). Higher education value added using multiple outcomes. *Journal of Educational Measurement, 53*(3), 368–400.
- Minaya, V., & Agasisti, T. (2019). Evaluating the stability of school performance estimates over time. *Fiscal Studies, 40*(3), 401–425.
- Morganstein, D., & Wasserstein, R. (2014). ASA Statement on value-added models. *Statistics and Public Policy, 1*(1), 108–110.
- National Research Council (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: The National Academies Press, <https://doi.org/10.17226/12820>.
- NCLB (2001). No Child Left Behind Act of 2001, P.L. 107–110, 20 U.S.C. § 6319 (2002).
- NWEA (2020). *2020 NWEA MAP growth normative data overview*. <https://teach.mapnwea.org/impl/MAPGrowthNormativeDataOverview.pdf>.
- Page, G. L., Martin, E. S., Orellana, J., & González, J. (2017). Exploring complete school effectiveness via quantile value added. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 180*(1), 315–340.
- Penuel, W. R., & Farrell, C. C. (2017). Practice partnerships and ESSA: A learning agenda for the coming decade. In Esther Quintero (Ed.), *Teaching in context: The social side of reform* (pp. 181–200). Cambridge, MA: Harvard Education Press.

- Quin, D., Heerde, J. A., & Toumbourou, J. W. (2018). Teacher support within an ecological model of adolescent development: Predictors of school engagement. *Journal of School Psychology, 69*, 1–15.
- Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy, 6*(1), 43–74.
- Roschelle, J., Feng, M., Gallagher, H., Murphy, R., Harris, C., Kamdar, D., & Trinidad, G. (2014). *Recruiting participants for large-scale random assignment experiments in school settings*. Menlo Park, CA: SRI International.
- Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics, 125*(1), 175–214.
- Rothstein, J. (2017). *Revisiting the impacts of teachers*. (Institute for Research on Labor and Employment Working Paper #101-17). <https://irle.berkeley.edu/files/2017/Revisiting-the-Impacts-of-Teachers.pdf>.
- Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M. E., & Roth, J., ... M. B. Resnick (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Education and Behavioral Statistics, 29*, 11–36.
- Thomas, S. (2001). Dimensions of secondary school effectiveness: Comparative analyses across regions. *School Effectiveness and School Improvement, 12*, 285–322.
- US Department of Education (2016). Non-regulatory guidance: Using evidence to strengthen education investments.
- WWC (2020). *What works clearinghouse standards handbook, version 4.1*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, <https://ies.ed.gov/ncee/wwc/handbooks>.
- Yan, B. (2017). Cycle-based budgeting and continuous improvement at Jefferson County Public Schools: Year 2 report. ERIC Document Reproduction Service No. ED577286.
- Yan, B., & Hollands, F. (2018). To fund or to defund: Making the hard decisions. *School Business Affairs*.
- Yu, B. (2013). Evaluation of alternative difference-in-differences methods. *Society for Research on Educational Effectiveness*.