Leadership and Policy in Schools

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/nlps20

A Comparison of Three Methods for Providing Local Evidence to Inform School and District Budget Decisions

F. M. Hollands, R. Shand, B. Yan, S. M. Leach, D. Dossett, F. Chang & Y. Pan

To cite this article: F. M. Hollands, R. Shand, B. Yan, S. M. Leach, D. Dossett, F. Chang & Y. Pan (2022): A Comparison of Three Methods for Providing Local Evidence to Inform School and District Budget Decisions, Leadership and Policy in Schools, DOI: 10.1080/15700763.2022.2131581

To link to this article: https://doi.org/10.1080/15700763.2022.2131581

Published online: 19 Oct 2022.

LEADERSHIP AND POLICY IN SCHOOLS
https://doi.org/10.1080/15700763.2022.2131581

Routledge
Taylor & Francis Group

A Comparison of Three Methods for Providing Local Evidence to Inform School and District Budget Decisions

F. M. Hollands[a], R. Shand[b], B. Yan[c], S. M. Leach[c], D. Dossett[c], F. Chang[c], and Y. Pan[d]

[a]Department of Education Policy and Social Analysis, Teachers College, Columbia University, New York, USA; [b]School of Education, American University, Washington, DC, USA; [c]Accountability, Research, & Systems Improvement Division, Jefferson County Public Schools, Louisville, Kentucky, USA; [d]Education Global Practice, World Bank, Washington, DC, USA

ABSTRACT

School and district leaders make annual decisions about investing their budgets in a multitude of educational programs. Policy directives set expectations for investing in programs that show evidence of improving student outcomes. However, evaluating many simultaneously-implemented programs under typical school operating conditions is challenging. We investigated three methods – cost-effectiveness analysis, program value-added analysis, and academic return on investment – to assess how each one fares against three criteria: rigor of methodology, difficulty of execution, and usability of results for decision-making. We apply each method to three programs implemented in a large, U.S. school district: Reading Recovery, Restorative Practices, and school nurses.

Introduction

Each year, school and district leaders make decisions about investing their budgets in a multitude of educational programs and practices. The task is presently magnified by large influxes of federal COVID-19 relief funding. Scholars have consistently documented a variety of factors that influence educators' decisions about programs (Asen et al., 2013; Hollands et al., 2019; Honig & Coburn, 2008; Weiss, 1977), but policy pressure has increased for investments in evidence-based programs and practices (e.g., Every Student Succeeds Act, 2015; National Commission on Excellence in Education, 1983; No Child Left Behind Act, 2002; U.S. Department of Education, 2016). The theory is that use of evidence in decision-making will lead to better student outcomes (Heinrich & Good, 2018). However, many programs currently in use by schools have not yet been subjected to rigorous study (Chang et al., 2021; Organisation for Economic Co-operation and Development, 2015). This leaves the responsibility to local education agencies to evaluate many of their own programs in order to comply with policy directives. In turn, this creates a need to identify research methods that are sufficiently rigorous yet feasible to execute on a timely basis. The methods must also provide information that is easily interpreted by the relevant decision-makers (Lortie-Forgues et al., 2021).

Traditional evaluation methods such as randomized controlled trials and quasi-experiments rarely produce results quickly enough to inform timely decisions about program continuation or discontinuation (Yan, 2018). In addition, locally-relevant research evidence is more likely to influence decision-makers than evidence from other contexts (Penuel & Farrell, 2017). Research methods such as academic return on investment (AROI; Frank & Hovey, 2014; GFOA, 2017; Levenson, 2012; Levenson et al., 2014) and program value-added analysis (VAA) (Meyer, 1997; National Research Council, 2010; Shand et al., 2022) have been proposed as solutions to provide timely and locally-relevant information about school programs, but have not been directly evaluated against more

CONTACT F. M. Hollands fmh7@tc.columbia.edu Teachers College, Columbia University, New York, NY 10027

rigorous methods such as cost-effectiveness analysis (CEA). All three methods, AROI, program VAA, and CEA, aim to estimate the effectiveness of programs relative to some other condition. In CEA and AROI, the treatment and control conditions are well defined, while in VAA the comparison condition is less explicit. Cost Analysis Standards Project (2021) recommends experimental or quasi-experimental methods for estimating effectiveness in CEA, but it can be executed using any method that reliably estimates impact and for which treatment contrast is clear. Proponents of AROI also recommend causal methods to estimate increases in student learning such that, in theory, the effectiveness methods used for AROI and CEA may be the same. However, for feasibility reasons, AROI is sometimes estimated by creating comparison groups using administrative data (e.g., Leach & Yan, 2021a; Smith & Knapp, 2019). The three methods differ more clearly in the approach to estimating costs: VAA has no cost component at all, AROI focuses on budgeted expenditures, and CEA estimates the opportunity costs of all resources used in implementing a program.

In this paper, we investigate the advantages and limitations of each method for informing school and district budget decisions by applying them to three programs implemented in a large, southern school district in the U.S.: Jefferson County Public Schools (JCPS), KY. Each program – Reading Recovery, Restorative Practices, and a school nurse program – represents a district priority and is a significant investment of resources for JCPS and many other school districts. Our goal was to address the question of whether analyses of program effectiveness using CEA, program VAA, or AROI render metrics that are timely and feasible to produce and use, yet rigorous enough to inform annual budget decisions.

We first describe the school district and how each of the three programs, Reading Recovery, Restorative Practices, and the school nurse program, is implemented in its schools. Subsequently, we describe each of the three methods, CEA, AROI, and program VAA. We explain how we applied them to analyze these programs and summarize our results. Because CEA and VAA are well-established methods in education research, we provide the most detail on AROI which is relatively new and lacks standardized procedures. We address how each method fares against three criteria: rigor of the methodology, difficulty of execution, and usability of the results for decision-making. In our discussion, we address the pros and cons of each method and their potential application to a broader range of programs. We also comment on how our findings compare with those produced by more traditional methods in published evaluations.

Setting and Context

JCPS is a large, urban school district located in Louisville, KY. With 96,304 students and 6,178 teachers in 168 schools, JCPS is the largest district in the state and one of the largest in the U.S. In 2020–21, the district's budget exceeded $1.6 billion, with per-pupil spending of more than $16,000. The district's school-based decision-making model grants broad authority to each school with regard to hiring, budgeting, and policymaking. School-based decision-making councils typically consist of a school administrator and a few elected teachers and parents. Over half of JCPS students belong to minority populations (37% African American, 13% Hispanic/Latinx, 10% Other). In 2020–21, more than 63% of students were eligible for free or reduced-price lunch and 5% were homeless. Roughly 9% of students were classified as English Learners and 13% required individualized education plans. Teachers were predominantly white (83%) and female (73%), with 82% of all teachers holding at least a master's degree, and nearly 7% holding national board certifications. Although state assessments and star rankings were temporarily suspended in 2020 due to the COVID-19 pandemic, JCPS received two out of five stars in the Kentucky Department of Education's 2018–19 accountability ratings. Based on state testing in that same year, student proficiency at the elementary, middle, and high school levels was 46%, 50%, and 37% respectively for reading, and 40%, 35%, and 31% for math. The average high school graduation rate was 84% and the dropout rate was 3%.

Cycle-based Budgeting at JCPS

To promote efficient use of financial resources, JCPS developed a cycle-based budgeting model for expenditures identified as *investment items* that are funded by the district office in response to requests for discretionary funding. 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. The cycle-based budgeting model consists of three core components: "alignment between investments and strategic priorities, upfront expectations for outcomes and timeline, and periodic review of the investments based on the timeline." (Yan, 2017, p. 4).

The district's Investment Tracking System (Leach & Yan, 2021a, 2021b) provides the underlying technical infrastructure for cycle-based budgeting by tracking investments from initial proposal to end-of-cycle review. Review periods vary from 1 to 5 years. The majority of investment items are personnel positions such as Success Coaches or interventionists, but they may also be programs or initiatives. For example, in 2017–18, in response to a principal's budget request, the district invested approximately $27,000 in a math intervention teacher at one of its elementary schools. This investment item aligned with two elements of the district's strategic mission: eliminate achievement, learning, and opportunity gaps; and personalize learning. It was expected to lead to improvements in student performance on state standardized math tests and was approved for a one-year cycle to be reviewed in 2018–19.

Choice of Programs to Which We Applied CEA, AROI, and Program VAA

The three programs we studied at JCPS were chosen to meet several criteria. First, they are all programs that are already or were on track to be widely used in the district. Reading Recovery has been offered in JCPS for more than 30 years, with consistent staffing and participation levels since 2012. With respect to nursing, the district was expanding the number of schools with full-time nurses and experimenting with different strategies including a pilot program in which contract nurses were hired in three schools instead of district or school-employed nurses. Restorative Practices is relatively new to JCPS but is being rolled out gradually to all schools. Second, the program directors needed to be receptive to the scrutiny to which the programs were being subjected, willing to participate in data collection, and open to considering the findings in future implementation plans. Third, the programs needed to have accessible sources of data within the district's existing systems that would allow us to determine which schools and students were being served and what amount of investment had been made by the district.

Data on all three programs were available in the Investment Tracking System. Student participation in Reading Recovery and other reading and math interventions was also documented in the *Intervention Tab* of the district's student information system, Infinite Campus. The Intervention Tab is used by the Kentucky Department of Education to collect statewide data on individual student academic and behavioral interventions. We describe each program below, including a brief summary of how they are implemented at JCPS and existing evidence on effectiveness and costs.

Description of the Programs Studied and How They are Implemented at JCPS

Reading Recovery

Reading Recovery is an early literacy program which aims to bring low-performing 1st-grade students to average literacy levels through an intensive, accelerated learning intervention (Clay, 1994). Eligible students are identified by scores on the Observation Survey of Early Literacy (hereafter, OS), a six-task measure of early literacy developed by Marie Clay, the creator of Reading Recovery (Clay, 2016). Students participate in daily, 30-minute one-on-one sessions with a trained Reading Recovery teacher, typically over 12 to 20 weeks (RRCNA, 2018). The program has served over 2.3 million students in almost a thousand U.S. school districts (International Data Evaluation Center, 2018). By reducing the

number of students experiencing severe literacy difficulties, Reading Recovery is expected to reduce the long-term financial costs to school districts serving these students (Clay, 1998). The program has been evaluated many times and the evidence has generally been positive for a variety of reading outcomes (see What Works Clearinghouse, 2013). For example, D'Agostino and Harmey's (2016) international meta-analysis of Reading Recovery found an overall effect size of 0.59. Sirinides et al. (2018) conducted a study of a national scale up of Reading Recovery and found positive effects on the Iowa Test of Basic Skills total reading score (Cohen's d = 0.37). Most recently, May et al. (2022) confirmed positive short-term effects of Reading Recovery but found negative long-term effects as measured using 3rd and 4th grade state standardized test scores. Costs of implementing Reading Recovery were estimated by Hollands et al. (2013) at $53,370 per school or $4,144 per student in 2010 dollars, and by Simon (2011) at $53,059 per school or $6,631 per student in 2007 dollars.[1]

During 2017–18, which is the focal year of the present study, Reading Recovery was active in 43 JCPS schools, with three teacher leaders and 66 Reading Recovery teachers serving a total of 560 students (Goodenough et al., 2018). Teacher leaders are responsible for providing one-year, on-site training for Reading Recovery teachers and for overall program implementation (Schmitt et al., 2005). JCPS Reading Recovery teachers typically serve four students during each school semester. Teacher leaders serve two students per semester in addition to their other duties. Occasionally, a third or fourth round of students is served, for example, mobile students or those who are chronically absent. For the first round, 1st-grade classroom teachers rank their students by literacy level during the first few weeks of the school year. The bottom 20% are administered the OS by a certified Reading Recovery teacher and the lowest-scoring students are assigned to the program until all slots at that school are filled. Second round assignment is conducted in a similar manner. Most of the district's Reading Recovery teachers are full-time teachers who conduct small-group literacy instruction when they are not working with program participants one-on-one. Students participating only in those small groups were not included in the present evaluation because they are extraneous to the main crux of the Reading Recovery program. However, we note that participation in these groups was common among treatment and control Reading Recovery students in the current study and there were significant overlaps in other reading interventions between the two groups.

Restorative Practices

Restorative Practices are a set of practices that educators have adapted from the restorative justice approach in the criminal justice system. They aim to prevent, reduce, and respond to student behavioral problems by building positive relationships, developing community, and managing conflict and tensions (International Institute for Restorative Practices (IIRP), 2018; Mallett, 2016; Thorsborne & Blood, 2013). The approach has been adopted by schools around the globe, with interest in the U.S. growing stronger in the wake of recent unrest related to racial inequities. Restorative Practices include victim-offender mediation conferences, group conferences, and various types of circles (Darling-Hammond et al., 2020). Some of these activities are intended to be preemptive by building better relationships among and between students and school staff, while others are designed to restore relationships after a harmful incident. Instead of imposing traditional punishments for behavioral offenses, victims and offenders are supported in settling on restorative actions such as agreeing to change specific behaviors or performing community service.

Despite the appeal of Restorative Practices as a non-punitive approach to school discipline, the evidence regarding its effects on student behavior, school climate, and academic outcomes is mixed (Darling-Hammond et al., 2020). Augustine et al. (2018) found reductions in the suspension rate for elementary school students, but not for middle school students, in Pittsburgh Public Schools after 2 years of exposure to a version of Restorative Practices. They also observed reduced disparities in suspension rates between African American and white students, and between low- and higher-income students. No effects were found on academic outcomes for elementary school students and negative effects were observed for middle school students. Acosta et al. (2019) found no significant improvements in school connectedness, school climate, peer relationships, developmental outcomes, or

victimization in middle schools in Maine that were implementing Restorative Practices. Students who reported experiencing higher levels of Restorative Practices reported better outcomes, but implementation was fairly weak across schools. Higher fidelity of implementation for Restorative Practices has been associated with greater improvements in the racial discipline gap (Gregory et al., 2016). Hollands et al. (2022) report national average costs of $57,370 per school and $139 per student ($114 per student using local KY prices) in 2018 prices to implement Restorative Practices, above and beyond Positive Behavioral Interventions and Supports (PBIS). More generally, resource constraints have been identified as an obstacle to implementation (Gray et al., 2017).

In an effort to improve student behavior, attendance, and school culture, JCPS implemented PBIS district-wide in 2013. PBIS is a framework for choosing and implementing evidence-based practices with an emphasis on using data to track student outcomes and a system-wide approach to training and provision of ongoing support for educators (Sugai & Simonsen, 2012). Despite evidence that PBIS has been effective in other contexts (Bradshaw et al., 2009, 2010), JCPS did not observe improvements in the target outcomes. Lack of results and an increasing desire to find ways to reduce racial disparities in student treatment led the district to integrate Restorative Practices with PBIS (Winsch, 2017).

Implementation of Restorative Practices began with the establishment of a 26-member District Leadership Team in 2016–17. These individuals participated in 11 days of training over the course of the year, delivered by trainers from the International Institute for Restorative Practices (IIRP). Thirty schools were selected to participate in the program over the next 3 years, with 10 schools being onboarded per year. The first year of program implementation involved three elementary schools, two middle schools, one high school, one combined middle/high school, and three special schools. Personnel from schools implementing Restorative Practices participated in initial half- or full-day training sessions and received ongoing support from IIRP trainers during monthly, day-long, on-site consultations (International Institute for Restorative Practices (IIRP), 2018). While the training and ongoing support represented additional time burdens for the personnel involved, no additional time was set aside for implementation with students because the approach was integrated with existing classroom activities. According to a local evaluation, implementation of Restorative Practices at JCPS in 2017–18 involved " . . . formulating 3–5 school-wide expectations, introducing circles into classrooms, speaking restoratively, using the 'with' style, and remembering to focus on the person, not the actions conducted by them." (Brahim et al., 2018, p. 7).

School Nurse Program

School nurses are expected to support educational achievement by improving student attendance. Their responsibilities include addressing student health problems, case-management services, and active collaboration with physicians (Magalnick et al., 2008). Reduced absenteeism and higher attendance are related to greater student achievement (Caldas, 1993; Gottfried, 2013; Lamdin, 1996; London et al., 2016; Moonie et al., 2008). By 2016, 82% of U.S. public schools employed a full- or part-time nurse (National Center for Education Statistics, 2020).

In recent years, evidence has emerged linking school nurses and the services they provide to better student outcomes. However, studies using causal methods are still lacking (Best et al., 2018). Wang et al. (2014) report that on-site school nursing services have been shown to improve student health and attendance, reduce early dismissals, and reduce teacher time spent on student health issues. Their cost-benefit analysis of school nurses in Massachusetts estimated the local cost per school of providing a full-time RN to be $84,724 in 2009 dollars and showed a return on investment (ROI) ranging from $0.7 to $3.8 for every dollar invested. Darnell et al. (2019) found that students in Kentucky high schools employing full-time nurses had higher graduation rates and lower absenteeism than students in schools with no nurse. In addition, students in schools with either a full-time or part-time nurse had higher ACT scores than students in schools without a nurse. Best et al. (2021) found that North Carolina students who suffered from asthma or diabetes were absent less often and earned higher grades when school nurses were responsible for fewer students. Similarly, Telljohann et al. (2004), studying a large, urban district in the Midwest, found that poor African American students with

asthma who attended schools with a full-time nurse missed fewer days than did their counterparts at schools with a part-time nurse.

In 2007–08, JCPS hired three Advanced Practice Registered Nurses (APRNs) to serve seven high-need elementary schools with the objective of targeting student health needs and improving academic achievement (Winsch, 2016). After a year of implementation, JCPS decided to provide Licensed Practical Nurses (LPNs) to additional schools. These LPNs were supervised by regional APRNs. The scope of practice for LPNs is narrower than for APRNs but is considered sufficient by the district for the majority of necessary duties within schools. By 2018–19, 35 of the district's schools were staffed by one or more full-time nurses, mostly LPNs, while 132 schools had no full-time nurse. All schools are required to provide basic health services provided by trained unlicensed assistive personnel. Full-time contracted LPNs in the three schools that are part of the current pilot program are assigned specific duties and must meet regular reporting requirements to the district health coordinator.

Methods

We evaluated Reading Recovery, Restorative Practices, and the school nurse program as implemented in JCPS schools using three different methods: CEA, program VAA, and AROI. JCPS established three criteria against which to evaluate the methods and findings: rigor of the methodology, difficulty of execution, and usability of the results for decision-making. We describe each method in turn and its application to the three programs, summarizing key details in Table 1. In our results section, we summarize the results of the analyses and address how each method fares against the district's criteria.

Cost-effectiveness Analysis

CEA juxtaposes the economic costs of a program against its effectiveness. CEA has been applied to education programs for over 50 years using the ingredients method to estimate costs (Levin, 1970, 1975; Levin & McEwan, 2001; Levin et al., 2018). This approach does not rely on budgets but, instead, identifies all resources utilized in the implementation of an intervention and values them based on their opportunity costs (Levin & McEwan, 2001). Resources include personnel time, materials and equipment, facilities, and other inputs such as trainer fees, travel expenses, and transportation. Cost-effectiveness ratios, calculated as incremental costs of a program per unit increase in the relevant outcome measure (above and beyond a control group), can be used to rank alternative programs in terms of efficiency. Decision-makers can use this information to determine which program will produce the greatest impact for a given investment (Kraft, 2020). While such analyses are potentially valuable to practitioners when choosing among programs, it is rare to find programs that have been evaluated using directly comparable measures, thereby limiting the currently available evidence on cost-effectiveness. Researchers have long identified tremendous potential for economic evaluation methods such as CEA to inform decision-making but have also documented the slow uptake of these methods in education in comparison to other human service and social science domains (Hummel-Rossi & Ashdown, 2002; Levin, 1988; Rice, 1997).

For each of the three programs studied, we designed a quasi-experiment to estimate the effects of the program as implemented in JCPS's schools. These are described individually below. To estimate costs of each of the three programs, we followed the ingredients method. We determined the type, quantity, and price of each ingredient used to implement Reading Recovery and Restorative Practices, and to provide nursing and health-related services in the schools included in our quasi-experiments. As recommended by Hollands et al. (2021) and Cost Analysis Standards Project (2021), we identified both local prices to inform local decision-makers and national average prices to provide results that could be more easily compared with other cost studies. We estimated average incremental costs per treatment school and per student compared to the control conditions.

Sources of data on types and quantities of resources used to implement each program included multiple interviews and e-mail exchanges with personnel responsible for supervising or implementing

Table 1. Summary of the three evaluation methods, CEA, VAA, and AROI, as applied to three programs.

Program	Method	Treatment group (T)	Comparison group (C)	Data requirement	Method for estimating results	Analysis level	Outcome measure	Method for estimating costs	Yr for effect data	Yr for cost data
Reading Recovery (RR)	CEA	First round RR students	Second round RR students	Pre and post outcome and demographic data for both RR groups Overlap between the two groups in outcome and control variables	QED: Cohort 1 vs. Cohort 2. Multilevel or OLS regression to predict Winter MAP RIT or OS scores from Fall MAP or OS and condition (i.e., T vs. C)	OS: Student MAP: Multilevel (teacher and student)	OS: Total Score MAP: Reading	Ingredients method to estimate economic costs for RR across the district and divided by total # of students served to obtain average cost per student	17–18	17–18
	VAA	RR students in Intervention Tab	Students receiving a hypothetical average effect reading intervention	Pre and post outcome and demographic data for students who participated in RR Pre and post outcome and demographic data for students participating in other reading interventions	OLS regression of 2018 scores on 2017 scores, with effect coding of interventions	Student	MAP Reading	N/A	17–18	N/A
	AROI	OS: First and second round RR students in 2 schools in Investment Tracking System (ITS) MAP: First round RR students in ITS	OS: Random sample of non-RR students MAP: Second round RR students	Pre and post outcome for both groups	Difference in differences (DiD) method	OS: Between-school subgroups MAP: Within-school subgroups	OS Total Score MAP Reading	Investment amount from budget for RR teacher compensation at each of 2 schools	17–18	17–18

(*Continued*)

Table 1. (Continued).

Program	Method	Treatment group (T)	Comparison group (C)	Data requirement	Method for estimating results	Analysis level	Outcome measure	Method for estimating costs	Yr for effect data	Yr for cost data
Restorative Practices (RP)	CEA	Students in 3 elementary and 3 middle schools that implemented RP in 2017–18	Students in 6 elementary and 2 middle schools that implemented RP in 2018–19	Pre and post outcome and demographic data for both groups	QED: Cohort 1 vs. Cohort 2. DiD regression (Quasi-Poisson distribution for referrals and suspensions)	Multilevel (school and student)	Referrals Suspensions School belonging Site safety Personal safety	Ingredients method to estimate incremental costs of RP for 6 T schools beyond costs of PBIS	17–18	Costs: 16–17 and 17–18
	VAA	One elementary school in the ITS data set that implemented RP in 2018–19	Elementary students in the ITS data set not directly involved in any end-of-cycle investment item	Pre and post outcome and demographic data for both groups	GLM regression (Quasi-Poisson distribution) of 2019 scores on 2018 scores	Student	Referrals Suspensions	N/A	18–19	N/A
	AROI	Same as CEA	Same as CEA	Pre and post outcome data for both groups	DiD method	Program (multiple schools, computed separately for elementary and middle)	Referrals Suspensions (per-student means)	Investment amount from budget for resource teacher compensation	17–18	17–18

(*Continued*)

Table 1. (Continued).

Program	Method	Treatment group (T)	Comparison group (C)	Data requirement	Method for estimating results	Analysis level	Outcome measure	Method for estimating costs	Yr for effect data	Yr for cost data
School Nurse	CEA	Students in 23 elementary schools with LPNs	Students in a matched set of 23 elementary schools without a nurse	Pre and post outcome and demographic data for both groups	QED: Optimal Multilevel Matching to create comparable groups of nurse (T) vs. no-nurse (C) schools DiD regression to compare T and C	School	Chronic absenteeism Attendance	Ingredients method to estimate economic costs of nursing at 23 T and 23 C schools	18–19	18–19
	VAA	Students in 7 out of 23 CEA T schools in the ITS data set with full-time LPN (LPN);Students in pilot contract LPN elementary school in ITS (Pilot)	Elementary students in the ITS data set not directly involved in any end-of-cycle investment item	Pre and post outcome and demographic data for both groups	OLS regression (attendance) and GLM regression (chronic absenteeism; binomial distribution) of 2019 scores on 2018 scores	Student	Chronic absenteeism Attendance	N/A	18–19	N/A
	AROI	Same as CEA (LPN) Same as VAA (Pilot)	Same as CEA (LPN)Same as VAA (Pilot)	Pre and post outcome data for both groups	DiD for LPN vs no-LPN schools (LPN) DiD for pilot school (Pilot)	Program (LPN) School (Pilot)	Chronic absenteeism Attendance	Investment amount from budget for nurse compensation	18–19	18–19

the programs at the JCPS district office, and local evaluation reports which included observation and fidelity of implementation data. In addition, we reviewed implementation guides, schedules of training dates, training rosters, lists of school supplies from JCPS schools and warehouses, invoices, JCPS travel guidelines, and district budget spreadsheets. We used Google Maps to estimate travel distances. Following the JCPS school schedule, we used a 174-day, 1,392-hour year for our cost calculations unless the district's schedule of workdays by position indicated otherwise for certain staff positions.

Sources of local price data included a public database of JCPS personnel salaries, JCPS schedules for salaries and number of workdays by position type, and JCPS's fringe rate calculator provided by the Human Resources department. National median or mean salaries and average fringe benefits rates were sourced from the Bureau of Labor Statistics; the School Superintendents Association; the American Association of University Professors; the National Center for Education Statistics; and the Organization for Economic Co-operation and Development. Many school personnel positions are not listed in national surveys, requiring judgment in selecting a suitable price for an equivalent position. Prices for general-use materials and equipment were found from national suppliers such as Best Buy, Staples, and CDW. Other price sources included Reading Recovery conference registration materials; university websites and the National Center for Education Statistics for college tuition fees; the Bureau of Transportation Statistics for airfares; and the U.S. General Services Administration for hotel and per diem rates. We used the inflation calculator provided by the Bureau of Labor Statistics, which is based on the Consumer Price Index, All Urban Consumers (CPI-U), to adjust all prices to 2017 for Reading Recovery and 2018 for Restorative Practices and the school nurse program in order to align with the timing of effectiveness data collected for each program.

Costs of physical space were calculated using Wang et al.'s (2020) Cost of Facilities Calculator. We assumed training took place in elementary school buildings and we amortized construction costs uprated by 21% for furnishing, fees, and equipment ("Living on Campus, 2011) over 30 years. We identified regional construction costs from Cumming Corporation, an international project management and cost consulting firm, or from School Planning & Management magazine. The closest regional school construction price we could identify for JCPS was for Raleigh Durham, NC. We calculated a national average price per square foot for school buildings by averaging school construction costs reported by Cumming Corporation for 20 U.S. regions.

We spread costs of introductory training and materials over 7 years on the basis that the benefits of these start-up inputs would last over this time. We chose 7 years because this was the average tenure of Reading Recovery teachers and, for comparability, we maintained this assumption for the other programs. We made an exception in the case of one school that dropped Restorative Practices at the end of one year. For this school, we attributed all introductory training and materials costs to the first year of implementation.

Reading Recovery CEA Study Design

Students participating in Reading Recovery at any JCPS school during the first semester of 2017–18 constituted the treatment group and those served in the second semester constituted the control group. This natural quasi-experimental design was used to compare winter NWEA Measures of Academic Progress (MAP) and OS scores between first (n = 272) and second (n = 249) round Reading Recovery students. Of the 521 total students who received Reading Recovery in the first (treatment group) or second (control group) round, a total of 471 students had both fall (pretest) and winter (posttest) MAP scores and a total of 316 students had both fall and winter OS scores. MAP scores were missing at random. OS scores were not missing at random, with only 30% of the control group having OS pretest scores. Therefore, an analytic OS sample (n = 133) was created by including only students whose Reading Recovery teachers had pretest and posttest results for 75% or more of their students. On average, the treatment group had lower pretest OS and Fall MAP scores than the control group. The OS results also hold for each individual school but the control group had lower pretest MAP scores than the treatment group in nearly one-third of the study schools (n = 13). We estimated an Ordinary Least Squares (OLS) regression model for OS scores and a multilevel (students

nested in teachers) regression model for MAP scores, each regressing treatment condition on Winter posttest scores while controlling for Fall pretest scores.

We estimated costs of delivering Reading Recovery for all students in the district and divided the total by the number of students served to obtain an average cost per student. This presumes costs are homogeneous across students. Students not participating in Reading Recovery benefited from a variety of reading instruction and supports, but these were also provided to the Reading Recovery students so we surmised that Reading Recovery supplements other reading practices at JCPS rather than replacing them. Accordingly, we assumed that all costs of Reading Recovery were incremental to the control group condition.

Restorative Practices CEA Study Design

To assess effectiveness of Restorative Practices, we took advantage of the staggered implementation of the program and compared outcomes for students in schools implementing the program in 2017–18 with outcomes for students in the schools that implemented the program in 2018–19. The treatment group consisted of 3 elementary and 3 middle schools which adopted Restorative Practices in 2017–18. We excluded the special schools and the one high school to improve comparability between the treatment and control schools. The treatment schools served a total of 2,477 students, averaging 413 students per school. The control group included 6 elementary and 2 middle schools which implemented Restorative Practices the following year, 2018–19. These 8 schools served 4,483 students, averaging 560 per school.

This design resulted in a treatment group demonstrating worse behavior issues to start with: a greater percentage of the middle school treatment students received referrals (63% vs. 28%) and suspensions (28% vs. 8%) and those students averaged more referrals (11.70 vs. 7.36) and suspensions (3.53 vs. 3.37) than their control group peers in the year prior to implementation. Among the elementary schools studied, a smaller percentage of treatment group students received referrals but those students averaged more referrals than their control group counterparts in 2016–17. In addition, a greater percentage of treatment group students received suspensions, but they averaged fewer suspensions than control group students in 2016–17. We estimated a separate difference-in-difference regression model with standard errors clustered at the school level for each of our five outcomes. Because referrals and suspensions were overdispersed count data (cf., Gage et al., 2016), their respective generalized linear models (GLMs) assumed a Quasi-Poisson distribution (e.g., Ver Hoef & Boveng, 2007). We used OLS regression for our three school survey outcomes (school belonging, site safety, and personal safety). Model specification is discussed in greater detail in Hollands et al. (2022).

All JCPS schools implement PBIS, so we treated costs of Restorative Practices as supplemental to the costs of implementing PBIS. We collected data on start-up costs to train a 26-person District Leadership Team in 2016–17 and approximately 420 school and district personnel in 2017–18, and ongoing costs for the first year of Restorative Practices implementation with 10 schools (2017–18). Ongoing costs of implementing the program were mostly related to monthly on-site coaching and support provided by district office personnel. We parsed out the costs for the 6 schools in our CEA and estimated costs per school and per student. Details of this analysis are reported in Hollands et al. (2022).

School Nurse Program CEA Study Design

To estimate effects of the school nurse program at JCPS, we used Optimal Multilevel Matching (OMM; Pimentel et al., 2018) to produce a set of highly comparable treatment and control elementary schools, the former with a full-time LPN and the latter without, and compared the two sets of schools. The pool of schools from which we drew our study sample included 25 elementary schools staffed by a full-time LPN and 59 elementary schools without a full-time nurse. Because we had clustered multilevel observational data, we used OMM to create the set of matched schools in order to approximate randomization (Pimentel et al.). The OMM algorithm utilizes unit- and cluster-level variables to

minimize the global distance between matched clusters and units. First, we computed a distance matrix for schools based on student-level characteristics: minority status (binary), gender, free or reduced-price lunch status, special education status, limited English proficiency, and prior year attendance. Although we did not match students 1:1, the OMM algorithm imposes a distance penalty for potential school-level matches that are not balanced on student-level characteristics according to the distance matrix.

Subsequently, the schools were matched on school-level covariates: percent of students at novice level in math, percent of students at novice level in reading, percent minority (nonwhite), need index score, prior year attendance, and prior year chronic absenteeism. We used the OMM fine balance algorithm in the R package MatchMulti which is ideal for small samples (Pimentel et al.). We were able to find close matches for 23 of the 25 schools with full-time LPNs. A difference-in-differences regression was performed to compare attendance and chronic absenteeism between the treatment and control group during 2018–19. We estimated costs of health services at both the treatment and control schools during the same year and subtracted the control school costs from the treatment school costs in order to obtain the incremental costs of the full-time LPN condition.

Program Value-added Analysis

VAA estimates the effect of a particular educational input, often teachers, on educational outcomes. Because students have different starting places and learn at different paces, and such differences are often related to the inputs they receive, simply focusing on raw differences in educational outcomes between students who have been exposed to different inputs is likely to be biased. VAA attempts to address this issue by focusing on growth, rather than achievement, and by statistically controlling for other factors, such as student and school characteristics, that could affect educational outcomes to isolate the effects of particular inputs (Koedel et al., 2015). VAA models can be specified in many different ways, but a common approach is to predict student expected academic performance based on prior performance and other factors and then estimate deviations of actual from predicted performance. The average deviation for a given input – a teacher, school, or program – is the value-added of that input to the educational production process. Although VAA has mostly been applied to teacher and school evaluation, Meyer (1997) and the National Research Council (2010) suggest that it can be applied to program evaluation. There have been a few such applications, primarily using teacher value-added as an outcome variable in determining the effectiveness of teacher preparation and professional development programs (Darling-Hammond et al., 2010; Harris & Sass, 2011). However, the benefits, challenges, and implications of measuring the value-added of individual interventions is relatively unexplored.

We applied a version of VAA to estimate the unique contribution of each of the three programs to student academic achievement while controlling for alternative interventions students may have also been participating in, or that comparison group students may have received. Details of these analyses are reported in Shand et al. (2022). We tested the application of the VAA method to evaluate the contribution of a wide range of JCPS interventions simultaneously using two data sets described in more detail below. Reading Recovery was among the interventions included in one data set. Investments in Restorative Practices and the school nurse program were in the second data set. The program VAA method only investigates the effectiveness of programs, with no attention to costs.

Reading Recovery VAA

Reading Recovery is among the interventions included in JCPS's Intervention Tab data set, which the district uses to track student-level participation in academic and behavioral interventions funded by certain state grants. This data set comprises over 70,000 student-intervention level records each year and, for a subset of students – generally early elementary students, students who are struggling academically, and students who are in struggling schools – includes a wide-ranging and nearly exhaustive list of interventions. Since the data set only includes students who are receiving targeted

interventions, this mitigates concerns about selection bias in two ways: Students are more nearly comparable with one another in the first place, and it is less likely that students in a comparison group for a given intervention are receiving other interventions that we cannot observe. For Reading Recovery, the comparison group is K-2 students receiving other targeted reading interventions.

We applied value-added regression models with effect coding (see, Mihaly et al., 2010) to the 2017–18 Intervention Tab data set to evaluate the relative contributions to reading outcomes of 18 early elementary literacy interventions, including Reading Recovery. In our preferred model, we regressed the 2018 MAP reading score standardized within grade level on the 2017 score, student characteristics, and a series of indicator variables for each intervention. We assessed improvements in MAP reading scores for 961 unique students participating in early elementary reading interventions. The estimated coefficient for each intervention variable represents the effect of that intervention relative to all other interventions. The VAA did not use OS scores as an outcome measure because it is only applicable to Reading Recovery and not to any of the other reading interventions in the analysis.

Restorative Practices and School Nurse Program VAA

Restorative Practices and the school nurse program are among the interventions included in the district's Investment Tracking System which supports JCPS's cycle-based budgeting initiative described earlier. Interventions may target all students in a school, a subgroup of students with common needs and characteristics, or a specific list of students. Our analysis focused on interventions that were slated for end-of-cycle budget review in 2018–19. To limit confounding due to comparison students receiving interventions we could not observe because they were funded from sources other than those included in the Investment Tracking System, or were simply not up for end-of-cycle review in 2018–19, we include elementary schools that had *any* active investments in the cycle-based budgeting system, even those not up for review in that particular year, but exclude schools with no active investment items. The comparison group is students in schools with targeted but no school-wide interventions who themselves were not targeted for intervention. A total of 8,551 students appear in the data set.

Because we have a comparison group that received no interventions to serve as an omitted category, for simplicity and to allow for separate estimation of the effects of different interventions for students participating in multiple interventions, interventions are dummy coded. Therefore, coefficients are interpreted relative to the omitted category of students in schools with targeted interventions who were not themselves the target of the interventions. Outcomes of interest include disciplinary referrals and suspensions for Restorative Practices and attendance and chronic absenteeism for nursing. The models are otherwise similar to those described for Reading Recovery, with the exception that we used nonlinear methods rather than OLS to estimate the effects on outcomes that are truncated, not continuous, or otherwise do not meet the functional form and distributional assumption requirements for OLS.

Academic Return on Investment

The introduction of AROI as a metric is an attempt to render the traditional methods of economic evaluation more practitioner-friendly and useful as a decision-making tool for annual budget allocation. It has been adapted from the business community's ROI metrics, primarily by educational consulting firms, as a way to help education administrators allocate budgets efficiently (Frank & Hovey, 2014; GFOA, 2017; Levenson, 2012). The typical AROI formula proposed by Levenson is as follows:

$$AROI = \frac{(Increase\ in\ Student\ Learning)\ x\ (Number\ of\ Students\ Helped)}{(\$\ Spent)}$$

The basic premise of AROI is that decisions about funding educational programs should consider both the budget amounts invested and student gains obtained as a result of implementing the program (note that a "program" may be a single investment item or several investment items targeting the same general outcome, which JCPS terms a "strategy"). The theory of change associated with AROI is that, if school and district administrators increasingly direct funds toward programs that, per dollar invested, yield the highest student gains, they will maximize the overall educational gains that can be obtained using the resources available. Program efficiency may also increase due to accountability pressure and to increased feedback as a result of the formative self-evaluation needed to calculate AROI.

To assess increase in student learning for an AROI analysis of a particular educational program, Levenson et al. (2014) recommend collecting "before and after" data (p. 11) and comparing results from one program with those of a control group or alternative program. Frank and Hovey (2014) suggest that education decision-makers refer to peer-reviewed research journals to predict the expected impact of a program that has been studied elsewhere in their own context, conduct their own impact studies, or, to estimate the effects of human capital policies, analyze teacher effectiveness data. Frank and Hovey extend the AROI concept to "System-Strategy ROI" which encourages school administrators to consider multiple options for addressing a particular student need and estimating the costs and estimated impacts on student outcomes for each one before choosing one to implement. For the cost component of AROI, Levenson et al. (2014) recommend adjusting the way budgets are reported to create a "program budget which collects all the costs (and only the costs) associated with a particular program." (p. 7).

Arithmetically, AROI is the inverse of a cost-effectiveness ratio (Yan, 2020), but JCPS's approach is different from how cost-effectiveness of a program is traditionally estimated and interpreted. On the effectiveness side, AROI does not aim to isolate the effect of the program under investigation from the effects of other programs simultaneously being implemented. Rather, the goal is to evaluate whether the program is effective at improving the intended outcome for the target population, either as a result of the program or its interaction with other existing programs. On the cost side, AROI includes only expended amounts on specific investment items related to a program, as opposed to economic costs of the resources used to implement the program, as used in CEA. For example, the investment amount used in AROI may be a Reading Recovery teacher's salary and fringe benefits, but the economic costs of implementing the program include the value of ongoing training, supervision by Teacher Leaders, use of materials, and physical space. In developing its AROI formulation, JCPS aims to balance rigor and practicality, leaving the metric open to criticism that it is neither rigorous nor practical.

JCPS's goals for investing its budget are typically stated at the level of one or a group of schools, or the entire district even if investments are targeted to subgroups. For example, the principal of ABC Elementary might decide to invest funds for 3 years in an interventionist to work with a subgroup of struggling readers, with a goal of reducing the school's percentage of novices on the annual state reading assessment. The goal here is school-wide, thus we consider the school as the relevant unit or cohort for AROI. During the 3-year investment cycle, the exact make-up of both the subgroup and ABC Elementary school as a whole will differ. Some students will remain in the subgroup during the cycle, thereby receiving multiple doses of treatment, whereas others will exit for various reasons. Similar churn will occur among students throughout the school. Despite changes in the composition of participants, the outcome measure remains static: state assessment data. While decision-makers might want to know if the interventionist was able to help subgroup members exit novice status, ultimately the success or failure of the investment will be judged by whether the overall percentage of novices at the school was reduced year-to-year. Such school-level results are generally the basis on which school leaders are evaluated and thus tend to drive their actions more than other types of data. For the majority of JCPS investments, the stated goal is at the cohort level rather than at the subgroup level, regardless of whether the cohort is a school, a group of schools, or the entire district. Because of its flexibility, AROI can be easily adapted for examining outcomes at subgroup and cohort levels with minor coding changes.

To estimate effectiveness for an AROI calculation, JCPS uses two methods to compare outcomes for the students benefiting from the investment item with outcomes for a comparison group: Previous cohort (PC) and difference in differences (DiD). The calculations are as follows:

Previous Cohort

$$\mathbf{AROI_{PC}} = \frac{WMA_{cycle} - WMA_{Baseline}}{Cost}$$

Cost = average annual budget cost/(number of students * $1,000)

WMA_{Cycle} = weighted moving average of outcome over investment cycle

$WMA_{Baseline}$ = weighted moving average of outcome for 3 years prior to investment start

Interpretation: change in outcome per $1,000/student relative to the unit prior to the investment (typically, the unit will be a school or subgroup).

Difference in Differences

$$\mathbf{AROI_{DID}} = \frac{\left(WMA_{cycle_{(T)}} - WMA_{Baseline(T)}\right) - \left(WMA_{cycle_{(c)}} - WMA_{Baseline(c)}\right)}{Cost}$$

Cost = same as above

$WMA_{Cycle(T)}$ = weighted moving average of treatment group outcome over investment cycle

$WMA_{Baseline(T)}$ = weighted moving average of treatment group outcome for 3 years prior to investment start

$WMA_{Cycle(C)}$ = weighted moving average of control group outcome over investment cycle

$WMA_{Baseline(C)}$ = weighted moving average of control group outcome for 3 years prior to investment start.

Interpretation: change in outcome per $1,000/student relative to similar units during the investment cycle (e.g., all high schools or all English learners or all first graders).

$$WMA = \frac{y_1 * n + y_2 * (n-1) + \ldots + y_n}{\frac{n(n+1)}{2}}$$

y = outcome of interest, n = time period.

The PC method compares student outcomes between the treatment group(s) and previous cohorts that share the same characteristics. The previous cohort may be the exact same group of students targeted for an intervention. In practice, the typical goal for investments is school- or district-wide (e.g., a reduction in reading novices or increased attendance) so, even when investments target specific subgroups, it is rarely the exact same set of students. The DiD method compares changes in outcomes between the treatment group(s) and the comparison group(s) during the investment cycle. When possible and appropriate, weighted moving averages are used to account for baseline and in-cycle trends. This is not always possible due to data availability or changes in reporting standards.

The choice of $1,000/student is arbitrary but does serve two purposes. First, from a practical reporting standpoint, AROI effects are often small relative to the budget expenditure so reporting in $1,000 increments brings the metric more in line with actual raw outcome values (e.g., test scores or chronic absenteeism rates). Second, dividing costs by the number of students accounts for implementation scope differences. Mathematically, this is the same as multiplying the average effect by the sample size, as advocated by Levenson (2012). Typically, the costs used for AROI have been the approved budget request proposal amounts, although many investments have much smaller actual expenditures. The implication of this discrepancy for AROI calculation and interpretation is discussed later. Similarly, unless specific student rosters are provided or indicated, the number of students served by an investment is assumed to be the number indicated in the budget request proposal as the intent-to-treat sample size. For school-wide investments, state-reported enrollment data are used.

Used together, the PC and DiD methods allow decision-makers to see the raw change in an outcome for a particular group (PC) while also attempting to account for overall trends (DiD). We use the term "treatment group" here because of the intentional flexibility of AROI to aggregate or disaggregate based on the scope of a particular investment. For example, the treatment group might be the literal treatment group for a single-school investment in a reading interventionist but we might calculate both school-level and aggregated district-level AROI metrics for a large-scale, multi-school reading improvement investment. Another example is the JCPS school nursing program which follows several different models: contract LPNs, school-purchased Registered Nurses, and a pilot nursing program in three schools. Here, a treatment group could be students in any combination of the nursing schools, from an individual LPN school to students in *any* school with some type of nurse. The aggregative flexibility of AROI means that the metric can be tailored to the intended use. This flexibility is also applicable to the relevant comparison group(s) for each version of AROI. District decision-makers may be more interested in fully aggregated district-level metrics whereas program personnel may find disaggregated school or subgroup AROI metrics helpful for improving implementation.

Challenges in Applying AROI

The ad hoc and piecemeal implementation over time of many district programs, for example, the LPNs, presents a challenge for AROI (and other program evaluation methods) due to different start dates for these investments. Newly implemented programs such as the recent pilot contract nurse program represent the ideal scenario for AROI (and program VAA). A related issue is that many investment items have been renewed such that the cycle in a particular proposal may not represent the entire implementation period of a specific investment. This is also true when an item is new in the Investment Tracking System only in terms of the funding stream (e.g., a position that was previously funded by a grant is now supported by the general budget). In these cases, calculating AROI based strictly on the current investment cycle is problematic in that the baseline values include the investment and are likely to bias the effect calculation.

Comparing two investments which start at vastly different times may, without warrant, penalize one or the other, depending on the long-term trends for a given outcome measure (Yan, 2020). For example, if the long-term overall trend for suspensions is strictly increasing between 2010 and 2020, then the baseline for a program which started in 2015 will necessarily be lower than the baseline for one implemented in 2018, assuming the two cohorts follow a similar trend, which is a likely scenario. Both investments will show a negative AROI using the PC method, but their DiD AROI metrics will be largely dependent on the linearity of their trend relative to the trend of the comparison group, which should more closely mirror the overall trend. This issue is exacerbated when reporting standards or measures change, as is common in education. Comparing assessments that start and end around the same time can help mitigate these issues for both the DiD and PC methods. In general, however, the reliability of PC method results is greatly diminished when investments span long time periods.

Our VAA and DiD AROI approaches are similar. In essence, AROI DiD effect computation methods are simplified (i.e., no control variables) versions of our VAA regression models, although using dummy versus effect coding in VAA models will result in more or less comparability between the interpretation of VAA and DiD AROI effects. A key point of divergence between AROI and VAA, however, is sample selection, especially for targeted (i.e., not schoolwide) interventions. VAA requires participation rosters to clearly identify treatment and control groups across multiple investments included in a given model. Somewhat paradoxically, a participation roster may be insufficient for computing AROI if it does not specify the characteristics of the students that would allow a comparison group to be chosen. For investments targeting a subgroup of students that do not fit in a clearly defined demographic category (e.g., where selection for participation is local and perhaps idiosyncratic), identifying an appropriate comparison group for AROI can be difficult. Thus, AROI is best-suited for investments that are new and target either all students or clearly identifiable subgroups (e.g., students scoring below proficient on state assessments). Sample differences between AROI and

VAA are summarized in Table 1, Columns 3 (Treatment group) and 4 (Comparison group). Notably, because we treated school nursing as a schoolwide intervention and all three pilot schools were included in the Investment Tracking System, the school nursing pilot program is the only case in our study where we were able to use the same sample for VAA and AROI.

Reading Recovery AROI

For two elementary schools investing in Reading Recovery teachers during 2017–18, we compared gains in MAP scores from fall to winter for first and second round Reading Recovery students. We also compared gains in OS scores between Reading Recovery participants and a random sample of non-Reading Recovery students across all JCPS Reading Recovery schools. Pretest scores necessary for the PC method were not available because the intervention targets 1st-grade students so both metrics were calculated using only the DiD method with a single baseline (i.e., no WMA). The AROI metric is calculated as reading outcome gains or losses per $1,000 invested in each Reading Recovery student. The invested amount was the Reading Recovery teacher's 2017–18 compensation requested in the Investment Tracking System, adjusted by .5 because the teacher spends only half of their time on the program. The Reading Recovery teachers were requested at the school level, therefore we computed separate AROI metrics for each school instead of a single, cohort-level metric.

Restorative Practices AROI

We computed four DiD AROI metrics for Restorative Practices based on the same treatment and control schools used in the CEA (i.e., first- vs. second-year implementers) and on the two primary outcomes: referrals and suspensions during 2017–18. Because AROI does not allow us to control for differences in outcomes at different school levels, we computed separate metrics for each outcome for elementary and middle schools. Similarly, because AROI is not able to accommodate count data, we modified the outcomes by computing the average referrals and suspensions per student, aggregated to the school level. Due to changes in reporting standards for referrals between 2016 and 2017, we did not utilize WMA for AROI calculations, and instead relied on a single pretest value to establish a baseline for each metric. The investment amount for Restorative Practices is for the compensation of three resource teachers for 2017–18. This amount was spread over the 10 schools implementing the program that year.

School Nurse Program AROI

To facilitate comparisons with CEA and program VAA, we computed DiD AROI metrics for two different nursing models: LPNs and the pilot nursing program. The pilot program represents an ideal scenario for AROI and program VAA because of its recent implementation whereas the ad hoc implementation of LPNs across multiple years means that AROI (and VAA) metrics necessarily include LPNs in the baseline calculations. To compare AROI results with CEA, we computed group-level DiD AROI metrics based on chronic absenteeism and attendance in 2018–19 for the 23 elementary schools with LPNs included in the CEA, and used the same matched schools for the comparison group. To compare AROI and VAA results for the pilot program, we computed DiD AROI metrics for chronic absenteeism and attendance for the elementary pilot school in the 2018–19 end-of-cycle investment VAA data set compared to non-nurse elementary schools in that same data set. Because the nursing analyses here focused on elementary schools, we utilized a single-year baseline to maximize sample size. Additionally, because both LPNs and the pilot nursing program were funded via a single budget request proposal that did not provide per-school personnel costs, we used an average per-school annual nurse compensation rate of $50,000 for AROI calculations for both models.

Results

We first briefly report the results of all three analyses by program, summarizing them in Table 2. Subsequently, we address how each method performed against JCPS's criteria: rigor of methodology,

Table 2. Summary of CEA, AROI, and VAA results for three programs.

Program	Method	Effects	Costs	Summary metric
Reading Recovery	CEA	OS Total Score: $B = 51.15$* MAP: $B = -1.00$	$6,621 per student	Cost-effectiveness ratio: $129 per one-point increase in OS score
	AROI	OS Total Score: +54.15, +44.46 points MAP: −1.25, −6.75 points	$4,204; $5,000 per student	Expenditure-effectiveness ratio: $78; $112 per one-point increase in OS score
	VAA	MAP: $B = -0.19$	N/A	N/A
Restorative Practices	CEA	Referrals: $B = -0.12$ Suspensions: $B = -0.18$ School belonging: $B = -0.02$ Site safety: $B = -0.09$ Personal safety: $B = -0.23$*	$114 per student	N/A
	AROI	Elementary school: Referrals: +0.12 Suspensions: +0.01 Middle school: Referrals: +1.53 Suspensions: +0.25	Elementary school: $62 per student Middle school: $40 per student	N/A
	VAA	Referrals: $B = -0.17$ Suspensions: $B = 0.14$	N/A	N/A
School Nurse	CEA	Chronic absenteeism: $B = 0.01$ Attendance: $B = -0.003$	$181 per student at LPN schools $81 per student at non-LPN schools	N/A
	AROI	LPN: Chronic absenteeism: +0.01 Attendance: −0.003 Pilot: Chronic absenteeism: +0.05 Attendance: −0.01	LPN: $111 per student Pilot: $128 per student	N/A
	VAA	LPN: Chronic absenteeism: $B = 0.53$* Attendance: −0.005* Pilot: Chronic absenteeism: $B = 1.15$* Attendance: −0.01*	N/A	N/A

difficulty of execution, and usability of results for decision-making, summarizing our observations in Table 3.

Reading Recovery Results

Reading Recovery Effects

The statistically significant *condition* regression coefficient ($\beta = 51.15$, $SE = 7.08$, $p < .001$) in our CEA OS model indicated that the treatment group, on average, scored 51 points higher on the posttest than the comparison group when controlling for pretest scores. This translated to an effect size (Cohen's d) of 0.5. For the students in the two schools for which AROI DiD metrics were calculated, the average OS gains beyond the comparison group gains were 54.15 and 44.46 points. As noted earlier, the VAA did not use the OS as an outcome measure. The *condition* regression coefficient in our CEA MAP model was negative ($\beta = -1.00$, $SE = 0.78$, $p = .20$) but not statistically significant. Students in the two

Table 3. Assessment of three evaluation methods, AROI, and VAA against decision-makers' Criteria.

Criteria	CEA	AROI	VAA
Rigor of methodology	*Effects*: QEDs provide straightforward treatment contrasts, accurately identify students who do and do not receive treatment, and reduce bias that could be introduced by confounding factors. *Costs*: Ingredients methods captures the value of all resources needed to implement the program. Identification of resource use in the counterfactual and treatment conditions provides insights into quantity and quality of services provided to students and clarifies treatment contrast.	*Effects*: Treatment contrast may not be clear, i.e., whether an alternative program is received by comparison students. In the absence of rosters, effects are estimated based on assumptions about which students are or are not served by the program. For ongoing programs, prior cohorts are likely to have been exposed to the treatment, thereby reducing contrast and potential for improvement. If calculated for a single school, confounds effects of school and personnel characteristics with program effects. *Costs*: Expenditures may capture only part of school's or district's total investment in a program. Actual expenditures may differ from planned expenditures. For investment items that serve multiple programs, attributing the full budget amount to each one overstates the cost.	*Effects*: Rigor is dependent on the characteristics of the dataset and whether included interventions represent all possible interventions received by the students. If students participate in multiple interventions, the effects of each one cannot be separated. Treatment contrast not always straightforward: comparison may be to average of other interventions that target the same outcome, or to students in schools receiving an intervention but who are not themselves targets of that intervention. Outcome on which impact is estimated may not be a key target outcome for all programs included in the analysis. If a program included in VAA involves only one staff member at one school, cannot isolate the effect of the role they play from the effect of their personal traits and experience, limiting generalizability. *Costs*: Not typically a component of VAA.
Difficulty of execution	Opportunities for QEDs are limited. When a program is mandated, need to include realistic alternative programs in the analysis as opposed to simply a "no-treatment" comparison. All programs to be compared must have a common outcome. Gathering accurate cost data is time-consuming, especially if counterfactual includes a variety of programs.	Relies on extant data and relatively straightforward analytical techniques. Can be difficult to identify comparison groups for investments that do not target a predefined demographic group. Data are readily available on expenditures although clarification may be needed on what percentage to attribute to the program.	Requires accurate data on which students participate in which interventions. Compiling existing data is time-consuming. Needs significant data analysis expertise. All programs to be compared must have a common outcome.

(*Continued*)

Table 3. (Continued).

Criteria	CEA	AROI	VAA
Usability of results for decision-making	*Pros*: Effectiveness results are reasonably credible and easy to interpret. Economic costs capture the overall burden on personnel time, materials and equipment, facilities, and other inputs. Helpful for cross-context comparisons, e.g., for different districts. Cost analysis provides valuable information about how programs are being implemented and can be improved. *Cons*: CEAs on many programs per year in order to inform annual budget decisions are not feasible. Economic costs do not relate directly to annual budget planning, unless expenditures are also presented. Spreading costs over multiple years obscures when funds are needed. *Best used for*: evaluating complex programs used at scale to inform improvement and equitable distribution of resources.	*Pros*: Timely for annual budget decisions. Reports expenditures directly relevant to budget planning. Flexibility in level at which effectiveness is measured provides information useful for different types of decision-maker. For programs operating for different lengths of time at different schools, AROI can be computed for each school, like replication studies. For a multi-school investment with a common implementation period across schools, school-level AROI metrics can be compared to identify uneven fidelity of implementation. *Cons*: Improvements hard to show for ongoing programs: may bias investments toward new programs. Best for comparing investments in programs that start and end at the same time. May bias investment toward students for whom it is easier to show large gains. Ignores burden on resources that are not identified with the program in the budget. May understate the ROI on investments where budgeted amount is underspent or where the item contributes to multiple programs. *Best used for*: evaluating discrete investments in personnel and programs to inform annual budget decisions.	*Pros*: Helps identify outlier programs: outperforming programs can be identified for adoption in more schools, underperforming programs can be targeted for improvement or phased out. Allows evaluation of multiple, simultaneously-implemented programs at once, reflecting the reality of school operation. *Cons*: Interpretation of effectiveness metric is not obvious. Programs included in the analysis which do not primarily target the common outcome measure may appear less effective. Inability to separate effects when students participate in multiple interventions reduces value of results for decisions about individual programs. Cost information or ROI metric not provided. *Best used for*: simultaneously evaluating a large number of discrete interventions to identify outliers.

schools for which AROI DiD metrics were calculated gained 1.25 points and 6.75 points less on MAP Reading than the comparison group. Students in the VAA sample whose only reading intervention was Reading Recovery scored lower (β = −0.19, *SE* = 0.17, p = .26) on MAP Reading than students participating in other reading interventions, although the result was not statistically significant.

Reading Recovery Costs

In our CEA of Reading Recovery, we estimated economic costs per student of $6,621 in local JCPS 2017 prices. The Reading Recovery teachers accounted for almost 80% of the economic costs. Budgeted expenditures on the Reading Recovery teacher's compensation at the schools for which AROI was calculated were $4,204 or $5,000 per student (equivalent to 63% and 76% of average economic costs per student respectively). These amounts differed substantially because one RR teacher was new and therefore received lower compensation. No cost estimate accompanied the VAA.

Reading Recovery Summary Metrics

Combining economic costs and effects for OS scores in the CEA, we obtain a cost-effectiveness ratio of $6,621/0.50 = $13,242 per *SD* increase in OS score. This can also be expressed as a cost per point increase in OS score beyond the comparison group of $6,621/51.15 = $129. In the AROI analysis, budget expenditures per point increase in OS score are $4,204/54.15 = $78 for one school and $5,000/44.46 = $112 for the second school. The AROI metric used at JCPS presents these as points per $1,000 spent per student, that is 12.88 (54.15/($4,204/$1,000)) or 8.89 (44.46/($5,000/$1,000)) points increase in OS score per $1,000 per student spent.

Using MAP as the outcome, we do not calculate a cost-effectiveness ratio because it appears that $6,621 in economic costs are incurred to obtain worse MAP outcomes. Similarly, in the AROI analysis, budget expenditures of $4,204 and $5,000 per student are incurred to obtain worse MAP outcomes. We do not compute AROI metrics when the effects are negative as there is clearly no ROI.

Restorative Practices Results

Restorative Practices Effects

Based on GLM regressions assuming a Quasi-Poisson distribution with standard errors clustered at the school level, our CEA found that implementation of Restorative Practices was associated with statistically insignificant reductions in the expected rates for referrals (11%, *SE* = 0.13, p = .35) and suspensions (17%, *SE* = 0.12, p = .12). Additionally, linear regression coefficients from school survey DiD models with clustered standard errors for school belonging (β = −0.02, *SE* = 0.06, p = .72), site safety (β = −0.09, *SE* = 0.10, p = .38), and personal safety (β = −0.23, *SE* = 0.10, p = .03) scores were all negative, but only the last was statistically different from zero. AROI DiD metrics based on the same schools included in the CEA showed increased mean referrals and suspensions per student in treatment elementary (+0.12, +0.01) and middle (+1.53, +0.25) schools relative to comparison schools. VAA regression coefficients for referrals (β = −0.17, *SE* = 0.25, p = .50) and suspensions (β = 0.14, *SE* = 0.53, p = .80) were not statistically significant.

Restorative Practices Costs

Incremental costs per student for a year of Restorative Practices implementation above and beyond PBIS in our CEA were $114 per student in 2018 local prices (see Hollands et al., 2022 for details). Compensation for the resource teachers accounted for 41% of the economic costs. Per student budget expenditures resulting from AROI analyses were $62 in elementary schools and $40 in middle schools, on average equal to 45% of economic costs. As noted earlier, this represents only the expenditures directly attributed to Restorative Practices in the Investment Tracking System, covering compensation for three full-time equivalent resource teachers spread across 10 schools. Information from CEA data collection led us to discover additional budgeted expenditures of approximately $108 per student for training and materials. These costs are included in the CEA but are treated as start-up costs and spread

over 7 years. We also discovered that, because JCPS resource teachers spread their time across multiple programs, Restorative Practices utilized 2.4 full-time equivalent resource teachers rather than three. No cost estimate is produced in VAA.

Restorative Practices Summary Metrics

Because our effects were all either statistically insignificant or in an undesired direction (i.e., more referrals and suspensions, or lower school climate and culture scores), we did not compute summary metrics for Restorative Practices as per usual CEA and AROI procedures.

School Nurse Program Results

School Nurse Program Effects

The results of our CEA showed that there were no statistically significant effects of a full-time LPN on student attendance ($\beta = -0.003$, $SE = 0.002$, $p = .25$) and chronic absenteeism ($\beta = 0.01$, $SE = 0.01$, $p = .29$) in the 23 matched pairs of schools (see Leach et al., in press). AROI DiD metrics for the same schools included in the CEA indicated a slight increase in the percent of chronically absent students (+1%) and slight decrease in attendance (−0.3%) in LPN schools relative to comparison schools with no full-time nurse. VAA results indicated a slight decrease in attendance (−0.45%, $SE = .001$, $p < .01$) and a slight increase in the odds of chronic absenteeism ($\text{Exp}(\beta) = 1.70$, $SE = 0.14$, $p < .001$) for elementary students in schools with LPNs compared to elementary students in the data set who were not targeted by any of the end-of-cycle investment items. VAA results for the elementary pilot nursing school showed a 1% decrease in attendance ($SE = .003$, $p < .001$) and increased odds of chronic absenteeism ($\text{Exp}(\beta) = 3.16$, $SE = 0.21$, $p < .001$) compared with the same control students as the LPN group. AROI DiD metrics for the elementary pilot nursing school showed an increase in chronic absenteeism (5%) and a decrease in attendance (−1%) relative to the same comparison schools in the VAA analysis.

School Nurse Program Costs

For the matched set of 23 schools, we found that health-related services provided in the control schools without a full-time LPN cost approximately $39,000 per school, or $81 per student, while treatment schools with a full-time LPN cost almost $82,000 per school, or $181 per student. The incremental costs of providing a school nurse program staffed by a full-time LPN were therefore $42,579 per school, or $100 per student. Treatment schools incurred additional costs due to the full-time nurse, extra materials and equipment, and a health office. Control schools had fewer health offices, lower personnel costs, and fewer materials. Annual compensation for the LPNs accounted for 59% of the economic costs in the schools staffed by a full-time LPN. Although effects were computed relative to comparison schools, the total (rather than incremental) per-student cost of $181 is more directly comparable to AROI budgeted expenditures per student because costs incurred by the counterfactual are not subtracted from the AROI cost metric. Per-student budget expenditures were $111 for the LPN program and $128 for the pilot program (respectively, equivalent to 61% and 71% of economic costs).

School Nurse Program Summary Metrics

Because our effects were all either statistically insignificant or in an undesired direction (i.e., higher chronic absenteeism or lower attendance), we did not compute a cost-effectiveness ratio or AROI metric for the school nurse program.

Evaluating the Three Methods against the School District's Criteria

Rigor of Methodology

CEA. Our quasi-experimental CEA methods varied in the extent to which they could eliminate sources of bias in the estimates of program effectiveness. The estimates for Reading Recovery and

Restorative Practices are based on a form of DiD analysis, which can provide a valid causal estimate but only under the relatively strong assumption that the changes over time for the control group are an appropriate counterfactual for what would have happened to the treatment group in the absence of treatment (Athey & Imbens, 2017). A limitation of our analysis is that we do not generally have multiple prior years of outcome data to establish a pre-treatment trend or formally test for parallel trends between treatment and control groups. However, both of our analyses rely upon staggered implementation, meaning that both treatment and control groups intend to receive the treatment. This reduces concerns about bias due to selection into treatment. If such staggering is random, then the DiD estimator provides an unbiased treatment effect (Athey & Imbens, 2018). Timing of intervention in our case is not random – the treatment students in Reading Recovery were slated to be treated first because they have observably lower pretreatment reading test scores – so there may still be some bias remaining, but focusing on a comparison group which will also ultimately receive the treatment does reduce this somewhat.

For the nursing intervention, we combined DiD with matching. Each of these methods has limitations. Matching can improve comparability between treatment and comparison groups on observable characteristics but cannot match on unobservables. DiD will net out time-invariant unobservables by focusing on changes over time but requires treatment and comparison groups to have similar trends. There is some evidence that combining them can help strengthen the overall approach whereby the strengths of one method compensate for the weaknesses of the other (Smith & Todd, 2005; Behrman et al., 2012). In the CEAs, the investment amount was based on opportunity costs of all resources used to implement the programs and therefore represented the most accurate reflection of resource use. Identification of resource use in the counterfactual as well as the treatment condition provides insights into quantity and quality of services provided to students and clarifies treatment contrast.

VAA. The rigor of value-added estimates depends on whether the observable student, school, and other characteristics, especially pre-intervention outcome measures, adequately control for unobserved student characteristics that are related to assignment to interventions. The methodological issues with VAA, including selection of appropriate covariates, whether to include student and school fixed or random effects, incorporation of multiple years of pre-treatment outcome data, and model specification, are discussed at length in Koedel et al. (2015), among other sources. Although there is some disagreement in the literature, experimental and quasi-experimental evidence suggests that VAA estimates of teacher and school effectiveness exhibit minimal bias on average, although individual estimates can be inaccurate and estimates tend to be quite noisy, especially in the middle of the effectiveness distribution with more accurate and precise measures in the upper and lower tails (Chetty et al., 2014; Kane et al., 2013). This, combined with the fact that VAA is inherently relative, suggests VAA is more useful for identifying outlier interventions that are performing particularly well or poorly for further study, replication, or discontinuation, rather than as a strict ranking of all interventions.

A key remaining question is whether research establishing the validity of VAA for teachers and schools extends to programs and interventions. Shand et al. (2022) examined this question by comparing VAA results to extant research that uses rigorous experimental and quasi-experimental methods and found that VAA results align reasonably well when data are available on a relatively exhaustive set of interventions targeted to specific subgroups of students. If students are participating differentially in interventions not captured in the dataset, the method is less useful. The interventions included in a program VAA should ideally all target the same outcome (e.g., improving reading performance) but, in practice, they are likely to address different aspects of learning. For example, some reading interventions may target vocabulary outcomes while others focus on overall comprehension. The result is that those interventions most aligned with the outcome measure (e.g., MAP Reading) will be at an advantage. Treatment contrast is not always straightforward: the comparison group may be the average of other interventions in the dataset that target the same outcome, or to students in the same dataset that are in schools receiving an intervention but are not themselves targets

of that intervention. Additionally, if students participate in multiple interventions, the effects of each one cannot be separated. A further limitation is that if a program included in a VAA involves only one staff member at one school, the effect of the role they play (e.g., interventionist) cannot be isolated from the effect of their personal traits (e.g., being social) and experience (e.g., working in the same school for many years), limiting generalizability of the finding.

AROI. The PC method of calculating AROI is essentially a single group quasi-experimental design. This is a relatively weak quasi-experimental method for making causal inferences (Shadish et al., 2002) because comparing outcomes against previous cohorts does not account for historical trends. The DiD method allows for trend comparisons. Although it is not always possible to compute both metrics for a given investment (as was the case for the PC method in all three programs described in this study), for investments with positive PC *and* DiD AROI metrics, we may conclude that, based on raw comparisons, the treatment cohort improved relative to previous similar cohorts *and* to their peers during the investment cycle. Where the PC and DiD AROI differ in direction, it is likely that a more detailed quantitative and qualitative analysis is required to help understand why the cohort trend was dissimilar to the overall trend.

However, for both DiD and PC AROI methods, it may be unclear whether the treatment or comparison group has previously or is currently implementing the same or similar programs or services using different sources of funding. This should be investigated in order to clarify the treatment contrast. In the absence of rosters indicating which students participate in which programs, effects are estimated based on assumptions about which students are or are not served by the program. In general, AROI is not designed to partial out individual investment effects. If, for example, a school simultaneously invests in a resource teacher, an interventionist, and an online program to help struggling math students, the AROI of each of these three investments is not independent of the others and should be combined as one strategy. The notion of an investment strategy is analogous to effect coding combinations of interventions for the Intervention Tab VAA.

In addition to reliance, in practice, on non-causal methods to estimate effectiveness, AROI suffers from a number of additional limitations. First, when the investment item is a single personnel position or in a single school, the type of investment is confounded with the individual hired to fill the position and/or with the school, making it hard to assess to what extent outcomes are influenced by the strategy versus the specific person or school implementing it. Second, the metric only includes dollars budgeted for the investment item rather than the full costs of implementing the strategy, which are likely to be substantially higher. Third, if an investment is made simply to continue funding an existing position or practice, it is unlikely that significant improvement will be observed. Simply maintaining the current performance level may be considered adequate. Fourth, many investments target multiple outcomes (e.g., academic and behavioral), which presents a challenge for assigning costs to each outcome. Currently, separate AROIs are computed for each outcome, with the total amount of the investment utilized for each metric. This may tend to bias the AROI metrics of multiple-outcome vs. single-outcome investments in nonsystematic ways. Alternatively, a single investment item may serve multiple programs, for example, when resource teachers spend time on each of several behavioral programs (e.g., Restorative Practices). Attributing the full budget amount to any one of these programs overstates the amount actually dedicated to it. As a result of these limitations, the AROI metric should be considered informational rather than definitive and should be accompanied by other qualitative and contextual information.

Difficulty of Execution

CEA. Effectiveness estimates for all three CEAs were obtained using existing data, primarily collected from Infinite Campus, with the Reading Recovery analysis also relying heavily on the International Data Evaluation Center data dump. However, without high levels of cooperation from program personnel in each of the three programs, the appropriate study designs and compilation of necessary data sets would not have been possible. Data compilation efforts were also facilitated by programmers

and subject-matter experts in the district's research department (e.g., assembling data, identifying reporting standards, coding budget request proposals). Opportunities for quasi-experimental designs (QEDs) using existing data are limited to situations where there is a comparable set of students or schools not participating in the same program. For services that are required or mandated by statutes, for example, interventions for struggling students or corrective improvement actions in schools, the comparison condition should provide an acceptable alternative service as opposed to being simply a "no-treatment" comparison. Comparative CEAs in which several alternative programs are investigated at once would provide more useful information but would be even more time-consuming.

Collecting data for the cost analyses on the types and quantities of resources used for program implementation was a time-consuming process involving multiple analysts, district personnel, and data sources. Each cost analysis took several months to complete and, as a result, we executed only three CEAs over approximately 18 months. The accuracy of our cost estimates was greatly enhanced by our insider access to information. However, we still had to gather and cross-reference a variety of sources of existing data. For example, for each individual trained, we had to determine whether they were certified or classified staff to determine fringe rates and stipend allowances; review training rosters to ascertain the numbers of hours in training; and consult the JCPS schedule of annual hours of work by position type, a salary database, and a fringe rate calculator to calculate the costs of each trainee's time. The amount of effort required is compounded when the counterfactual includes a variety of programs or services.

Ideally, CEA results for each program would be compared with those from other similar interventions to allow decision-makers to assess which one provides the best ROI. With only three programs studied to date and each one targeting different student needs, the effort needed to conduct CEA on the many programs implemented in a district appears unmanageable. Most districts cannot afford to hire analysts with the skill set needed to conduct rigorous evaluations and cost analyses, or to contract with external researchers for this purpose.

VAA. Program VAA is an efficient method of comparing multiple programs in one analysis when suitable data sets are available, but compiling the data initially is a substantial task. For example, care must be taken to include programs that target the same outcome and to correctly identify which students participate in which interventions. We were able to conduct the VAA for Reading Recovery, Restorative Practices, and the school nurse program using existing administrative data from the Intervention Tab and Investment Tracking System data sets, although some additional confirmatory research with program leaders was necessary to better identify targeted subgroups of students for some interventions in the Investment Tracking System data set.

The data requirements for VAA are extensive and districts may not collect student-level participation data for a comprehensive set of interventions. However, when such data do exist, running value-added models is relatively straightforward by analysts with specialized training in regression analysis, including panel data with complex nesting structures (observations over time nested within students, who are, in turn, nested in interventions and schools) and specialized statistical software packages such as R or Stata. This expertise may be beyond what is available in smaller districts and require partnership across districts, or with research partners such as universities or educational consultancies. Setting up the data for analysis requires reshaping and coding decisions, such as effect coding versus dummy coding and how to treat students who are in more than one intervention. Once the data have been set up and models run once, subsequent data setup and analysis are easier. VAA does not require cost data. This information could be gathered separately in a similar manner to CEA or AROI and linked to the VAA effectiveness estimates but it would be daunting to conduct rigorous cost analysis for the many programs included in a single VAA. Additionally, there are challenges in assigning costs when the estimated effect is produced by a combination of programs or when a school staff member implements multiple intervention programs.

AROI. Among the three methods, AROI is the easiest to execute in terms of data collection, assembly, and computation. It does not involve sophisticated statistical controls and hypothesis testing and can be executed using spreadsheet software such as Microsoft Excel or calculators. Because of its simplicity, AROI analysis can be conducted annually by practitioners for a large number of programs. This is clearly an advantage over CEA and VAA. It circumvents the challenging task of isolating the independent effect of a new investment or its interaction effect with other variables, which simplifies the estimation and makes it feasible for broad adoption by practitioners. Clearly, this simplification can lead to bias and faulty decisions. For AROI involving analysis at the cohort level, the required multiple years of data are usually readily available since they are typically used for accountability purposes. Particular attention must be paid to the measurement of the outcomes of interest as reporting standards and assessments frequently change. AROI analysis that requires student participation data poses the same data collection challenge as CEA and VAA. In addition, identifying a suitable comparison group is difficult if the treatment group is not a clearly defined demographic category. For cost data, AROI is much simpler than the rigorous cost analysis in CEA as it simply relies on planned budget expenditures recorded in an existing system. However, clarification may need to be sought from program or personnel supervisors regarding what percentage of investment amount to attribute to the program in question.

Usability of the Results for Decision-making

CEA. The CEA effectiveness results were straightforward to interpret as the outcomes were measured using familiar metrics such as attendance, absenteeism, referrals, and test scores. In addition, because CEA requires that costs and effects are compared between well-defined treatment and control groups, the comparison group and treatment contrast were clear: LPN vs. no LPN, Reading Recovery round 1 students vs. round 2 students, and Restorative Practices integrated with PBIS vs. PBIS only. The cost analysis results, while a more rigorous assessment of resource use than that produced in AROI, were less well received by school and district decision-makers because of their practical need to focus on budgets. The concept of economic costs was deemed somewhat theoretical and unrelated to day-to-day decision-making which is more often about how to tweak and improve programs rather than making decisions about whether to adopt or end them. Certain items are considered as given operational expenses, so the primary concern for decision-makers is additional or saved budget line items, that is, an expenditure analysis as distinct from a cost analysis (see, Hollands et al., 2021). Furthermore, spreading start-up costs and durable items over multiple years obscures when funds are needed to acquire the relevant items. Economic costs may be of more value to decision-makers in other contexts who are considering adoption of the program and need to assess the feasibility of implementing the program in their own setting.

Despite decision-makers' relative disinterest in economic cost estimates (as opposed to actual expenditures) of the programs studied, the in-depth study of program implementation necessary in CEA to identify and value the resources used led to useful practical insights that informed continuous improvement efforts. For example, it became apparent that there were more Reading Recovery teachers than needed to serve struggling readers at some schools and fewer than needed at other schools, highlighting an inequitable distribution of resources. The cost analysis of Restorative Practices highlighted the fact that a key element of the program, restorative conferences, were not being implemented at all. Furthermore, the presentation of start-up costs broken out from recurring costs of the program demonstrated the wisdom of asking schools to commit to the program for multiple years in order to make the heavy initial investment in training worthwhile. Our cost analysis of the nursing program led the district health coordinator to revisit the program's theory of change and to develop a manual and training sessions to clarify nurses' expected activities.

VAA. The greatest appeal of VAA for decision-making purposes is that it allows evaluation of multiple, simultaneously-implemented programs at once, reflecting the reality of school operation. However, any program included in the analysis which does not primarily target the common outcome

measure may appear less effective than if it were independently evaluated for impact on the outcome it is specifically designed to improve. Value-added models are inherently relative and thus the interpretation of the results for a single program is not straightforward. For example, effectiveness may be relative to the average effect of a combination of other interventions. Unless the average effect can be benchmarked against national norms, it is hard to assess the extent to which any of the interventions are helping students perform at expected levels. Therefore, VAA models seem better suited to producing a relative ranking of programs to identify the best and worst-performing programs for further study than for evaluating a single program. Outperforming programs can be considered for adoption in more schools, while underperforming programs can be targeted for improvement or possibly phased out.

VAA models are often not transparent about the implied comparison, and interpretation of results can depend on how interventions are coded and what, if any, is an omitted or baseline category. If students receive more than one program in the data set, the effects of the two programs cannot be separated for that student, thereby limiting the value of results for decisions about each program. Further, VAA models do not include cost or financial data, precluding an ROI analysis.

AROI. The most valuable aspect of AROI is that it can be calculated for any academic program on a timely basis to inform annual budget decisions. Conceptually, the AROI results might be most relatable to practitioners because they provide information that is more closely aligned with the decision-making context of school and district administrators. In practice, schools frequently roll out multiple concurrent changes which may directly or indirectly impact the targeted student population. Practitioners need information about whether a new investment is effective when other changes are taking place rather than when everything else is held constant.

Because program personnel set the goals at the outset of the investment cycle, the return is estimated at the relevant cohort level (i.e., school, district, or subgroup) using metrics with which they are familiar and for which they are held accountable, for example, the percentage of students meeting standards on the state assessments. This flexibility in the level at which effectiveness is measured (student, subgroup of students, school, or district) provides information that is useful for different types of decision-maker (program implementer, principal, program director). There might be some value in comparing DiD metrics among multiple implementing subgroups within a larger cohort (e.g., for elementary vs. middle schools who implemented Restorative Practices) but, in general, we expect that decision-makers are more interested in learning how the overall strategy performed relative to alternative strategies (e.g., all elementary schools with a full-time LPN vs. all elementary schools without a full-time nurse). For investments that reach multiple schools over a common implementation period, school-level AROI metrics can be compared to help identify potential issues with fidelity of implementation. When investigating effects for programs that have been operating for different lengths of time at different schools, AROI can be computed for each one to include multiple data points for a single program. These data points can be viewed as replication studies contributing to a big picture view of program performance.

In spite of the potential appeal of AROI, the method's limitations must be acknowledged when using AROI metrics to inform decisions. Notably, improvements in effects may be hard to produce or detect for programs that have been operating for some years, potentially biasing investments toward new programs. Consequently, AROI seems best-suited for comparing investments with similar start dates and similar cycle lengths. Secondly, certain groups of students are naturally likely to show greater gains than others, such that programs serving these students will appear to have a higher AROI. Hill et al. (2008) demonstrated that effect sizes for educational outcomes decrease as students age. Consequently, comparing the AROI of programs serving fourth graders to those serving twelfth graders would most likely suggest that all funds should be redirected toward the youngest students. Programs that serve more challenged students may appear to provide lower ROI than those serving the general student population.

The AROI cost calculation uses budget amounts which relate directly to the annual financial decisions school administrators must make, rather than economic costs which might inform a grander consideration of resource allocation which is rarely, if ever, undertaken in schools. Thus, practitioners can interpret AROI results directly in terms of financial decisions they are required to make and metrics against which their schools are evaluated. However, focusing only on expenditures ignores the burden on resources that are not specifically identified in the budget as contributing to the program, and an overstatement of ROI. This could lead to decisions that result in an overcommitment of time for existing personnel. For investment items that contribute to multiple programs, the returns may be underestimated if the full expenditure amount is attributed to only one. AROI may also be underestimated in cases where the budget is underspent or where the comparison condition receives an alternative service whose expenditures should be subtracted from those of the treatment.

Discussion

Our investigation of three school programs using three different methods, program VAA, AROI, and CEA, produced results which were surprisingly consistent across methods in their findings regarding the effects of the three programs (or lack thereof), but disappointing in terms of ROI for the school district. For Reading Recovery, both CEA and AROI results showed positive effects based on the proximal measure, the OS, and a negative result based on the distal measure, MAP Reading. VAA results indicated that Reading Recovery is neither more nor less effective than the average reading intervention implemented at JCPS. All three sets of results for Restorative Practices indicated no significant improvements to student referrals and suspensions. The CEA also found no impact on two measures of school climate: school belonging and site safety, and a negative impact on personal safety. For the school nurse program, effects on attendance and chronic absenteeism were null for the CEA, and negative for both AROI and VAA. Overall, our effectiveness findings suggest that AROI and VAA methods produce results which are directionally similar to more traditional but time-consuming QEDs used for CEA. The fact that all three program results tend to agree across methods and samples is somewhat promising for further exploring the viability of sacrificing some rigor for practicality.

In contrast to the similarities across effectiveness metrics, we found that average economic costs per student were consistently higher than average budget expenditures per student: $2,000 higher for Reading Recovery (44% higher); $63 higher for Restorative Practices (124% higher); and $61.50 higher for the school nurse program (51% higher). This is despite the fact that, in the economic cost analysis, costs for training, materials, and durable items were spread over multiple years. The budgeted expenditures only cover compensation for certain personnel positions, which, while representing the largest cost component in each analysis, only account for 40–80% of the costs. Specifically, resource teachers accounted for 41% of the economic costs while LPNs accounted for 59% of the school nurse program costs, and the Reading Recovery teachers accounted for 80% of the economic costs. These differences reflect the fact that each program requires time from additional personnel beyond those for whom a budget expenditure is identified, in addition to materials, equipment, physical space, and other inputs. The difference is particularly large for Restorative Practices because the program involves training the entire staff at a school (both certified and classified positions). Additionally, implementation of Restorative Practices with students relies on regular teachers and administrators who are budgeted under their main positions. For the school nurse program, budgeted expenditures did not include compensation for other staff responsible for delivering health services in JCPS schools such as APRNs, contract nurses, the district health coordinator, and unlicensed assistive personnel.

In general, the more a program relies on implementation by personnel who must reallocate their time from responsibilities covered by budget amounts not directly tied to the program being evaluated, the greater will be the difference between the budgeted expenditure and economic costs. The danger in this discrepancy is that programs that appear to be low cost in terms of budget expenditures may require substantial personnel time that is not recognized. Implementing multiple “low cost” programs

of this nature could result in overburdening school staff and reducing the likelihood that any of the programs are implemented with fidelity and produce the desired outcomes.

Mostly null results precluded the calculation of cost-effectiveness ratios for the majority of our outcomes but, in the case of Reading Recovery, the economic costs to produce a one-point increase in OS scores ($129) are higher than the comparable AROI metrics ($78 and $112), reflecting greater economic costs. Despite the fact that Reading Recovery expenditures and economic costs demonstrated the smallest percentage difference, this discrepancy in the ROI metric illustrates how reliance on budget expenditures will likely result in an overstatement of the returns to an investment. However, in the absence of comparable AROI metrics or cost-effectiveness ratios from evaluations of the impact of alternative reading programs on OS scores, it is not clear how decision-makers should judge whether the observed returns are worth the investment. Because MAP Reading scores are closely tracked by decision-makers at the district and national benchmarks are available, it may be easier to assess whether a given improvement in MAP Reading scores is worthwhile. Clearly, at $6,621 per student in economic costs, or an average of $4,602 per student in budget expenditures, Reading Recovery does not appear to be a good investment to improve student performance on MAP Reading.

When effectiveness results are null for a program, decision-makers need to consider other benefits that may justify the costs, such as impact on other desirable outcomes, lack of alternatives to provide a mandated service, or fulfilling other decision-making criteria, for example, improving equity. At $114 per student in local prices, Restorative Practices appears relatively inexpensive and its appeal to local values may justify its continuation with more attention to fidelity of implementation. However, because the program serves all students in a school, the cost per student is not directly comparable to the costs of Reading Recovery. Total program costs should also be considered. In addition, these costs are above and beyond the costs of implementing PBIS. The nursing program is similarly low cost per student but school-wide. Because the non-LPN schools also incurred substantial costs for health services, the incremental cost metric from the CEA is similar to the budgeted amount. However, a fairer comparison is between the total costs per student of the LPN program ($181) and the budgeted expenditures for the LPNs. In the shadow of the COVID-19 pandemic, it is unlikely that schools will forego nurses, regardless of the evidence to support their contribution to student outcomes.

Comparison of Our Results with Existing Studies

In general, our effectiveness findings were either aligned with or more negative than existing studies which have used experimental or quasi-experimental methods. For Reading Recovery, our CEA's positive effect size of 0.50 for the OS score is lower than the 0.79 reported by D'Agostino and Harmey (2016) and 0.99 reported by Sirinides et al. (2018). However, our results are consistent with D'Agostino and Harmey's observation that effect sizes for Reading Recovery are larger when the OS is the outcome measure than when other measures are used, and that results are stronger for experimental studies than for quasi-experimental studies. MAP is not typically used to assess outcomes from Reading Recovery so we do not have a direct comparison for our negative CEA result, but a local evaluation by a school district in Wisconsin reported that former Reading Recovery students severely lagged the district average in meeting proficiency on MAP Reading in Grades 3–8 (5% vs. 38%; Madison Metropolitan School District, 2014) suggesting that the program is not well aligned with MAP or that the positive effects wear off over time, as found by May et al. (2022). It is also likely that our research design biased the results because the second round of Reading Recovery students (the control group) performed better than the treatment group on OS pretest scores in all schools but worse on Fall MAP scores in one third of the schools we analyzed. Our CEA cost estimate using national average prices, $7,325 per student ($6,621 is the estimate using local prices for the same resources), is in line with that of Simon (2011) which equates to $7,850 per student in 2017 dollars.

Our CEA of school nursing may be the most rigorous study to date of the costs of school nurses and their effects on chronic absenteeism and attendance for the general school population; there are no extant causal studies against which to assess our effectiveness methods and results. However, using less

rigorous methods, Wang et al. (2014) reported improvements in attendance and a positive ROI for school nursing programs. Darnell et al. (2019) and Best et al. (2021) also found positive associations between school nurse programs and attendance. A comparison of our cost results against previously published results adjusted to national equivalent 2018 prices shows that our estimate of approximately $116,000 per school using national average prices is substantially higher than Wang et al.'s (2014) adjusted estimate of $89,700 and Baisch et al.'s (2011) adjusted estimate of $96,900 (see Leach et al., Forthcoming, for details on adjustment). We would expect our estimate to be higher as neither of the others was based on the ingredients method and therefore most likely omitted some resources: Baisch et al.'s estimate only included the nurses' compensation while Wang et al. added costs of medical supplies.

Causal studies of Restorative Practices have not documented impact on referrals. Augustine et al. (2018) showed improvements in suspensions for elementary school students but not for middle school students. This aligns moderately well with our AROI results which are substantially worse for middle school students. Our negative (albeit nonsignificant for two of the three constructs) CEA results for student perceptions of school climate are generally consistent with findings by Augustine et al. and by Acosta et al. (2019). Other estimates of the costs of Restorative Practices are not available against which to compare ours.

Overall, a comparison of our results against existing evidence suggests that school districts cannot assume applicability of external evidence to their own contexts, consistent with broader concerns about generalizability (e.g., Hedges, 2018). This may be especially true when the external evidence comes from efficacy studies in which program implementation is supported by program developers or the research team. Additionally, adapting programs to reduce the burden on staff time and other resources may be counterproductive in that the local results are less positive than the results of the external study. However, if districts rely only on internal evidence using natural quasi-experiments, they may produce disappointing findings if treatment schools are consistently those most needing the intervention. Published estimates of costs, unless based on the ingredients method, may be closer to budget expenditures as opposed to reflecting the value of all resources required to implement a program. Careful consideration should be given to the burden of adopting the program on existing personnel, space, and equipment.

Conclusion

Our comparison of three methods of evaluating school programs as implemented under typical conditions provides promising evidence that program VAA and AROI can feasibly produce credible evidence on program effects to inform school budget decisions in a more timely and less burdensome manner than CEA. However, VAA provides no information on ROI, and AROI appears to undervalue the resources needed for program implementation. Combining CEA's more rigorous approach to estimating costs with AROI's more feasible methods for estimating program effects may represent a useful synthesis of methods for evaluating district programs to help inform decisions about whether to continue, discontinue, or scale up a program. VAA can usefully serve to evaluate many programs at once in order to identify outlier interventions that merit closer scrutiny.

Overall, these metrics can provide key pieces of evidence to support investment strategies for the district but should not be the only factors considered. Further applications are needed to a wider range of programs to confirm these initial results. In addition, it would be useful in future work to compare VAA and AROI using the exact same data set to calculate the metrics for multiple interventions. An ideal data set for comparing the two methods might be one in which each student only participates in one intervention, the students do not participate in any interventions the prior year, and the students are all in the same school the prior year.

Our findings suggest that the quality of local evidence used to inform annual decisions about re-investing in existing programs could be improved by the establishment of data systems and processes that accurately and exhaustively document student participation in individual programs implemented

in a school or district, and also the resources needed to implement them, particularly personnel time. While this would place a high data collection and entry burden on school staff, it would greatly enhance the ability to evaluate the impact of local programs on student outcomes and their costs. We expect that most, if not all, schools and districts already document participation in at least some programs and services such as those funded by the federal government, but often in different data systems. However, comprehensive evaluation of an education agency's programs requires exhaustive and integrated documentation of student participation in all programs.

Note

1. The large difference in cost per student arises because Simon (2011) assumes one Reading Recovery teacher per school, serving eight students, while Hollands et al. (2013) used the national average statistics for Reading Recovery of 1.6 teachers and 12.9 students per school.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

The research reported here was supported by theInstitute 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.

References

Acosta, J., Chinman, M., Ebener, P., Malone, P. S., Phillips, A., & Wilks, A. (2019). Evaluation of a whole-school change intervention: Findings from a two-year cluster-randomized trial of the restorative practices intervention. *Journal of Youth and Adolescence*, *48*(5), 876–890. https://doi.org/10.1007/s10964-019-01013-2

Asen, R., Gurke, D., Conners, P., Solomon, R., & Gumm, E. (2013). Research evidence and school board deliberations: Lessons from three Wisconsin school districts. *Educational Policy*, *27*(1), 33–63. https://doi.org/10.1177/0895904811429291

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32. https://doi.org/10.1257/jep.31.2.3

Athey, S., & Imbens, G. W. (2018). *Design-based analysis in difference-in-differences settings with staggered adoption*. National Bureau of Economic Research. No. w24963

Augustine, C. H., Engberg, J., Grimm, G. E., Lee, E., Wang, E. L., Christianson, K., & Joseph, A. A. (2018). *Can restorative practices improve school climate and curb suspensions? An evaluation of the impact of restorative practices in a mid-sized urban school district*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2840.html

Baisch, M. J., Lundeen, S. P., & Murphy, M. K. (2011). Evidence-based research on the value of school nurses in an urban school system. *Journal of School Health*, *81*(2), 74–80. https://doi.org/10.1111/j.1746-1561.2010.00563.x

Behrman, J. R., Gallardo-Garcia, J., Parker, S. W., Todd, P. E., & Vélez-Grajales, V. (2012). Are conditional cash transfers effective in urban areas? Evidence from Mexico. *Education Economics*, *20*(3), 233–259. https://doi.org/10.1080/09645292.2012.672792

Best, N. C., Nichols, A. O., Waller, A. E., Zomorodi, M., Pierre-Louis, B., Oppewal, S., & Travers, D. (2021). Impact of school nurse ratios and health services on selected student health and education outcomes: North Carolina, 2011-2016. *Journal of School Health*, *91*(6), 473–481. https://doi.org/10.1111/josh.13025

Best, N. C., Oppewal, S., & Travers, D. (2018). Exploring school nurse interventions and health and education outcomes: An integrative review. *Journal of School Nursing*, *34*(1), 14–27. https://doi.org/10.1177/1059840517745359

Bradshaw, C. P., Koth, C. W., Thornton, L. A., & Leaf, P. J. (2009). Altering school climate through school-wide positive behavioral interventions and supports: Findings from a group-randomized effectiveness trial. *Prevention Science*, *10*(2), 100–115. https://doi.org/10.1007/s11121-008-0114-9

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions*, *12*(3), 133–148. https://doi.org/10.1177/1098300709334798

Brahim, N., Hensel, S., & Carpenter, K. (2018). *Restorative Practices: 2017-2018 evaluation report*. JCPS Department of Accountability, Research, and Systems Improvement.

Caldas, S. J. (1993). Reexamination of input and process factor effects on public school achievement. *The Journal of Educational Research*, *86*(4), 206–214. https://doi.org/10.1080/00220671.1993.9941832

Chang, Y., Hollands, F. M., Holmes, V. R., Shand, R., Evans, P., Blodgett, R., Wang, Y., & Head, L. (2021, April 9–12). *Challenges to finding evidence for ESSA: A case study of southern urban district* [Paper presentation]. American Educational Research Association Annual Meeting.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593

Clay, M. M. (1994). Reading recovery: The wider implications of an educational innovation. *Literacy, Teaching and Learning*, *1*, 121–141. https://www.proquest.com/openview/d1d98b40c1b3892705c69661a5131714/1?pq-origsite=gscholar&cbl=39178

Clay, M. M. (1998). Accommodating diversity in early literacy learning. In D. R. Olson & N. Torrance (Eds.), *The handbook of education and human development: New models of learning, teaching and schooling* (pp. 197–217). Wiley-Blackwell.

Clay, M. M. (2016). *An observation survey of early literacy achievement* (3rd ed.). Heinemann.

Cost Analysis Standards Project. (2021). *Standards for the economic evaluation of educational and social programs*. American Institutes for Research. https://www.air.org/sites/default/files/Standards-for-the-Economic-Evaluation-of-Educational-and-Social-Programs-CASP-May-2021.pdf

D'Agostino, J. V., & Harmey, S. J. (2016). An international meta-analysis of Reading Recovery. *Journal of Education for Students Placed at Risk (JESPAR)*, *21*(1), 29–46. https://doi.org/10.1080/10824669.2015.1112746

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. https://doi.org/10.1080/02607476.2010.513844

Darling-Hammond, S., Fronius, T. A., Sutherland, H., Guckenburg, S., Petrosino, A., & Hurley, N. (2020). Effectiveness of restorative justice in U.S. K-12 schools: A review of quantitative research. *Contemporary School Psychology*, *24*(3), 295–308. https://doi.org/10.1007/s40688-020-00290-0

Darnell, T., Hager, K., & Loprinzi, P. D. (2019). The impact of school nurses in Kentucky public high schools. *The Journal of School Nursing*, *35*(6), 434–441. https://doi.org/10.1177/1059840518785954

Every Student Succeeds Act, 20 U.S.C. § 6301. (2015). https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Frank, S., & Hovey, D. (2014). *Return on investment in education: A "system-strategy" approach*. Education Resource Strategies (ERS). https://www.erstrategies.org/library/return_on_investment_in_education

Gage, N. A., Larson, A., Sugai, G., & Chafouleas, S. M. (2016). Student perceptions of school climate as predictors of office discipline referrals. *American Educational Research Journal*, *53*(3), 492–515. https://doi.org/10.3102/0002831216637349

GFOA. (2017). *Academic return on investment: Foundations and smart practices*. Government Finance Officers Association. https://gfoaorg.cdn.prismic.io/gfoaorg/142dffae-80a3-4878-8c52-6da0a87d42eb_AROIWhite+PaperFINAL.pdf

Goodenough, A., Henry, A., & Cicchiello-Wright, H. (2018). *Reading Recovery 2017-18 summary report*. Jefferson County Public Schools.

Gottfried, M. A. (2013). Quantifying the consequences of missing school: Linking school nurses to student absences to standardized achievement. *Teachers College Record*, *115*(6), 1–30. https://doi.org/10.1177/016146811311500605

Gray, A. M., Sirinides, P. M., Fink, R., Flack, A., DuBois, T., Morrison, K., & Hill, K. (2017). *Discipline in context: Suspension, climate, and PBIS in the school district of Philadelphia*. CPRE Research Reports.

Gregory, A., Clawson, K., Davis, A., & Gerewitz, J. (2016). The promise of restorative practices to transform teacher-student relationships and achieve equity in school discipline. *Journal of Educational and Psychological Consultation*, *26*(4), 325–353. https://doi.org/10.1080/10474412.2014.929950

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *957*(7–8), 798–812. https://doi.org/10.1016/j.jpubeco.2010.11.009

Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, *11*(1), 1–21. https://doi.org/10.1080/19345747.2017.1375583

Heinrich, C. J., & Good, A. (2018). Research-informed practice improvements: Exploring linkages between school district use of research evidence and educational outcomes over time. *School Effectiveness and School Improvement*, *29*(3), 418–445. https://doi.org/10.1080/09243453.2018.1445116

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Hollands, F., Pan, Y., & Escueta, M. (2019). What is the potential for applying cost-utility analysis to facilitate evidence-based decision-making in schools? *Educational Researcher*, *48*(5), 287–295. https://doi.org/10.3102/0013189X19852101

Hollands, F. M., Leach, S. M., Shand, R., Head, L., Wang, Y., Dossett, D., Hensel, S., . . . Hensel, S. (2022). Restorative Practices: Using local evidence on costs and student outcomes to inform school district decisions about behavioral interventions. *Journal of School Psychology*, *92*, 188–208. https://doi.org/10.1016/j.jsp.2022.03.007

Hollands, F. M., Pan, Y., Shand, R., Cheng, H., Levin, H. M., Belfield, C. R., Kieffer, M., Bowden, A. B., & Hanisch-Cerda, B. (2013). *Improving early literacy: Cost-effectiveness analysis of effective reading programs*. Teachers College, Columbia University.

Hollands, F. M., Pratt-Williams, J., & Shand, R. (2021). *Cost analysis standards & guidelines 1.1*. Cost Analysis in Practice (CAP) Project. https://capproject.org/resources

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. https://doi.org/10.1177/0895904807307067

Hummel-Rossi, B., & Ashdown, J. (2002). The state of cost-benefit and cost-effectiveness analyses in education. *Review of Educational Research*, *72*(1), 1–30. https://doi.org/10.3102/00346543072001001

International Data Evaluation Center. (2018). *Reading Recovery by the numbers*. https://readingrecovery.osu.edu/learn/Reading_Recovery_Facts_Figures.pdf

International Institute for Restorative Practices (IIRP). (2018). *Behavior Support Systems Model: Restorative Practices implementation guide*. https://www.jefferson.kyschools.us/sites/default/files/JCPS%20Restorative%20Practice%20Implementation%20Guide.pdf

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*. Measures of Effective Teaching Project. Bill & Melinda Gates Foundation. https://files.eric.ed.gov/fulltext/ED540959.pdf

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics and Education Review*, *47*, 180–195. https://doi.org/10.1016/j.econedurev.2015.01.006

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Lamdin, D. J. (1996). Evidence of student attendance as an independent variable in education production functions. *The Journal of Educational Research*, *89*(3), 155–162. https://doi.org/10.1080/00220671.1996.9941321

Leach, S. M., Hollands, F. M., Stone, E., Shand, R., Head, L., Wang, Y., Pan, Y., . . . Pan, Y. (in press). Costs and effects of school-based licensed practical nurses on elementary student attendance and chronic absenteeism. *Prevention Science*.

Leach, S. M., & Yan, B. (2021a). *Academic return-on-investment (AROI) and budget decision-making: A research brief*. Louisville, KY: Jefferson County Public Schools. https://www.jefferson.kyschools.us/sites/default/files/IES_AROI_Brief_3_Final_Dec_2021.pdf

Leach, S. M., & Yan, B. (2021b). *ITS 2.0 user's guide. Getting more bang for your buck: Academic return on investment and the Investment Tracking System. Accountability, research, & systems improvement*. Louisville, KY: Jefferson County Public Schools. https://www.jefferson.kyschools.us/sites/default/files/AROI_Brief_1_ARSI_Final.pdf

Levenson, N. (2012). *Smarter budgets, smarter schools: How to survive and thrive in tight times*. Harvard Education Press.

Levenson, N., Baehr, K., Smith, J. C., & Sullivan, C. (2014). *Spending money wisely: Getting the most from school district budgets*. District Management Council.

Levin, H. M. (1970). A cost-effectiveness analysis of teacher selection. *The Journal of Human Resources*, *5*(1), 24–33. https://doi.org/10.2307/144622

Levin, H. M. (1975). Cost-effectiveness analysis in evaluation research. In M. Guttentag & E. L. Struening (Eds.), *Handbook of evaluation research* (Vol. 2, pp. 89–122). Sage Publications.

Levin, H. M. (1988). Cost-effectiveness and educational policy. *Educational Evaluation and Policy Analysis*, *10*(1), 51–69. https://doi.org/10.3102/01623737010001051

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd) ed.). Sage Publications.

Levin, H. M., McEwan, P. J., Belfield, C., Bowden, A. B., & Shand, R. (2018). *Economic evaluation in education: Cost-effectiveness and benefit-cost analysis*. Sage Publications.

Living on Campus. (2011, June). *College Planning and Management Magazine*. 1105 Media.

London, R. A., Sanchez, M., & Castrechini, S. (2016). The dynamics of chronic absence and student achievement. *Education Policy Analysis Archives*, *24*(112). http://dx.doi.org/10.14507/epaa.24.2741112

Lortie-Forgues, H., Sio, U. N., & Inglis, M. (2021). How should educational effects be communicated to teachers? *Educational Researcher*, *50*(6), 345–354. https://doi.org/10.3102/0013189X20987856

Madison Metropolitan School District. (2014). 1*Reading Recovery evaluation*. https://accountability.madison.k12.wi.us/files/accountability/2014-11-1%20-%20Reading%20Recovery%20Evaluation.pdf

Magalnick, H., & Mazyck, D., American Academy of Pediatrics Council on School Health. (2008). Role of the school nurse in providing school health services. *Pediatrics*, *121*(5), 1052–1056. https://doi.org/10.1542/peds.2008-0382

Mallett, C. A. (2016). Truancy: It's not about skipping school. *Child and Adolescent Social Work Journal*, *33*(4), 337–347. https://doi.org/10.1007/s10560-015-0433-1

May, H., Blakeney, A., Shrestha, P., Mazal, M., & Kennedy, N. (2022). Long-term impacts of Reading Recovery through third and fourth grade: A regression discontinuity study from 2011-12 through 2016-17. [Paper presentation]. American Educational Research Association Annual Meeting, San Diego, CA.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, *16*(3), 283–301. https://doi.org/10.1016/S0272-7757(96)00081-7

Mihaly, K., McCaffrey, D. F., Lockwood, J. R., & Sass, T. R. (2010). Centering and reference groups for estimates of fixed effects: Modifications to felsdvreg. *The Stata Journal*, *10*(1), 82–103. https://doi.org/10.1177/1536867X1001000109

Moonie, S., Sterling, D. A., Figgs, L. W., & Castro, M. (2008). The relationship between school absence, academic performance, and asthma status. *Journal of School Health*, *78*(3), 140–148. https://doi.org/10.1111/j.1746-1561.2007.00276.x

National Center for Education Statistics. (2020). *School nurses in U.S. public schools. Data Point*. NCES 2020–086.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. *The Elementary School Journal*, *84*(2), 113–130. https://doi.org/10.1086/461348

National Research Council. (2010). *Getting value out of value-added: Report of a workshop*. The National Academies Press.

No Child Left Behind Act of 2001, P.L. 107–110, 20 U.S.C. § 6319. (2002).

Organisation for Economic Co-operation and Development. (2015) . *Education policy outlook 2015: Making reforms happen*. OECD Publishing.

Penuel, W. R., & Farrell, C. C. (2017). Practice partnerships and ESSA: A learning agenda for the coming decade. In E. Quintero (Ed.), *Teaching in context: The social side of education reform* (pp. 181–200). Harvard Education Press.

Pimentel, S. D., Page, L. C., Lenard, M., & Keele, L. (2018). Optimal multilevel matching using network flows: An application to a summer reading intervention. *The Annals of Applied Statistics*, *12*(3), 1479–1505. https://doi.org/10.1214/17-AOAS1118

Rice, J. K. (1997). Cost analysis in education: Paradox and possibility. *Educational Evaluation and Policy Analysis*, *19*(4), 309–317. https://doi.org/10.3102/01623737019004309

RRCNA. (2018). *Standards and guidelines of reading recovery in the United States* (8th ed.). Reading Recovery Council of North America.

Schmitt, M. C., Askew, B. J., Fountas, I. C., Lyons, C. A., & Pinnell, G. S. (2005). *Changing futures: The influence of Reading Recovery in the United States*. Reading Recovery Council of North America.

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

Shand, R., Leach, S., Hollands, F. M., Chang, F., Pan, Y., Yan, B., Head, L., . . . Head, L. (2022). Program value-added: A feasible method for providing evidence on the effectiveness of multiple programs implemented simultaneously in schools. *American Journal of Evaluation*. https://doi.org/10.1177/10982140211071017

Simon, J. (2011). *A cost-effectiveness analysis of early literacy interventions*. (Publication No. 865292034) [Doctoral dissertation, Columbia University]. ProQuest Dissertations and Theses Global.

Sirinides, P., Gray, A., & May, H. (2018). The impacts of reading recovery at scale: Results from the 4-year i3 external evaluation. *Educational Evaluation and Policy Analysis*, *40*(3), 316–335. https://doi.org/10.3102/0162373718764828

Smith, R., & Knapp, K. (2019). Return on Instructional Investment (ROII) model: A practical guide for school leaders. *Academy of Educational Leadership Journal*, *23*(1), 1–11. https://www.proquest.com/docview/2238974843?pq-origsite=gscholar&fromopenview=true

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, *125*(1–2), 305–353. https://doi.org/10.1016/j.jeconom.2004.04.011

Sugai, G., & Simonsen, B. (2012). *Positive behavioral interventions and supports: History, defining features, and misconceptions*. Center for PBIS & Center for Positive Behavioral Interventions and Supports, University of Connecticut. https://assets-global.website-files.com/5d3725188825e071f1670246/5d82be96e8178d30ae613263_pbis_revisited_june19r_2012.pdf

Telljohann, S. K., Dake, J. A., & Price, J. H. (2004). Effect of full-time versus part-time school nurses on attendance of elementary students with asthma. *The Journal of School Nursing*, *20*(6), 331–334. https://doi.org/10.1177/10598405040200060701

Thorsborne, M., & Blood, P. (2013). *Implementing restorative practices in schools: A practical guide to transforming school communities*. Jessica Kingsley Publishers.

U.S. Department of Education. (2016). *Non-regulatory guidance: Using evidence to strengthen education investments*. https://ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf

Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, *88*(11), 2766–2772. https://doi.org/10.1890/07-0043.1

Wang, Y., Hollands, F. M., Shand, R., Pratt-Williams, J., Head, L., Kushner, A., & Chang, Y. (2020). *Cost of facilities calculator*. [Computer app]. Cost Analysis in Practice (CAP) Project. https://capproject.org/resources

Wang, L. Y., Vernon-Smiley, M., Gapinski, M. A., Desisto, M., Maughan, E., & Sheetz, A. (2014). Cost-benefit study of school nursing services. *JAMA Pediatrics*, *168*(7), 642–648. https://doi.org/10.1001/jamapediatrics.2013.5441

Weiss, C. H. (1977). Research for policy's sake: The enlightenment function of social research. *Policy Analysis*, *3*(4), 531–545. https://www.jstor.org/stable/42783234

What Works Clearinghouse. (2013, July). *Beginning reading intervention report: Reading Recovery®*. http://whatworks.ed.gov

Winsch, B. (2017). *Restorative Practices evaluation plan*. Jefferson County Public Schools, KY.
Winsch, B. J. (2016). *School nurse program evaluation*.Department of Data Management, Planning, and Program Evaluation. Jefferson County Public Schools.
Yan, B. (2017). *Lessons learned from the budget requests and approvals: 2017–18*. Jefferson County Public Schools.
Yan, B. (2018, January). Improve system deficiencies to make stronger budget decisions. *School Business Affairs*, *84*(1), 14–17.
Yan, B. (2020). *Issues around using academic return on investment (A-ROI) for informing and improving decisions. Part I: Validity*. Jefferson County Public Schools, KY. https://eric.ed.gov/?id=ED603587