

# Achieving Better Educational Practices Through Research Evidence: A Critical Analysis and Case Illustration of Benefits and Challenges

ECNU Review of Education  
2021, Vol. 4(1) 108–127  
© The Author(s) 2020  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/2096531120916742  
journals.sagepub.com/home/roe



**Steven M. Ross**

Johns Hopkins University

**Jennifer R. Morrison** 

Johns Hopkins University

## Abstract

**Purpose:** We examine considerations regarding the positive contributions of evidence accountability and challenges that frustrate educators in gaining access to the needed product information.

**Design/Approach/Methods:** We review the research literature on the multiple characteristics of evidence relative to consumer (practitioner) interests. We then examine, through a “case illustration” of an initiative in a large school district, a second challenge for evidence usage—conducting viable studies and interpreting outcomes from comprehensive interventions in complex educational systems.

**Findings:** Despite attention being given to rigorous evidence, consumers report preferring peer recommendations and local pilot studies as sources. In our case illustration, we found that the availability of evidence from comprehensive formative evaluation studies was viewed by stakeholders as positively contributing to program implementation quality and sustainability over time.

---

## Corresponding author:

Jennifer R. Morrison, Center for Research and Reform in Education, School of Education, Johns Hopkins University, Baltimore, MD 21286, USA.

Email: [JRMorrison@jhu.edu](mailto:JRMorrison@jhu.edu)



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

**Originality/Value:** We use a real-world “case illustration” of a complex initiative in a large, diverse school district to illustrate how current policies and expectations regarding evidence support for educational programs is filtered through multiple agendas and personal needs of key stakeholders. Consequently, evaluators acquire nontraditional roles that go beyond routine execution of rigorous studies. Given these factors, we offer recommendations for fostering more meaningful and objective interpretations and usage of evidence by local stakeholders.

### Keywords

Evidence of effectiveness, program evaluation, technology integration

Date received: 21 November 2019; accepted: 12 March 2020

### Introduction

In recent years, federal policies in the U.S. have emphasized the importance of schools adopting programs and practices supported by evidence from rigorous research studies. Almost two decades ago, the No Child Left Behind Act (NCLB) of 2001 (U.S. Congress, 2001) explicitly identified a preference for randomized experiments in its definition of “scientifically based” research. In reviewing policies directed to bridging the gap between educational research and school practice, Farley-Ripple et al. (2018) note the subsequent influences of the Educational Sciences Reform Act of 2002 and the Institute of Educational Sciences (IES) whose mission is to disseminate to educators, researchers, and the public scientific evidence on educational practices. To advance this goal, the IES, in turn, created the What Works Clearinghouse (WWC) in 2002, which identifies and reviews rigorous studies that yield evidence evaluated as meeting standards *without reservations* or *with reservations* (IES, 2017). Rigorous research approaches employing randomized controlled trials (RCTs) or quasi-experimental designs (QEDs) are required to achieve these respective levels of strength.

More recently, the U.S. Department of Education has enacted a new major policy, the “Every Student Succeeds Act” (ESSA, 2015). One focus was reducing the extreme accountability emphasis on schools’ success in raising test scores and of federally applied punitive consequences of associated failures. A second focus was establishing and promoting standards of research evidence for evaluating educational programs for schools. ESSA defines four ordered tiers of evidence support. Relative to the NCLB era, this system is proving much more consequential to developers, practitioners, and researchers as a consequence of formally being applied by states and school districts for vetting and approving programs. Specifically, ESSA’s four tiers include (1) strong evidence (RCT), (2) moderate evidence (QED), (3) promising evidence (correlational study with statistical controls for selection bias), and (4) demonstration of a rationale (well-specified logic model informed by research).

As evaluators of educational programs, we have experienced firsthand the impacts of this burgeoning evidence movement through more active interests by both program vendors (developers and providers) and end users (school districts and schools) for studies that meet high standards of rigor such as ESSA Tiers 1 and 2. In the present article, we address these developments by first examining considerations regarding the positive contributions of evidence accountability as well as prevailing challenges that frustrate end-users' gaining access to the product information they need. Second, we present as a case illustration of the complexities of evidence gathering and interpretation of an examination of our recent experiences in evaluating a district-wide reform initiative focusing on technology integration in classroom instruction (Morrison et al., 2018).

### *Contributions and growing pains of the evidence dissemination*

Clearly, the evidence movement brings many important benefits to educational research and practice. On the positive side, there appears to be elevated interest by practitioners in identifying and purchasing educational programs backed by credible research evidence than was the case in the past (Morrison, Ross, & Cheung, 2019). Consumer access to research evidence has substantially expanded through continuing WWC reviews (<https://ies.ed.gov/ncee/wwc/>) and the more recently created Evidence for ESSA website (<https://www.evidenceforessa.org/>) housed at Johns Hopkins University. States and school districts increasingly are implementing vetting processes that approve funding only for interventions having adequate evidence support in accord with ESSA standards. For researchers, natural benefits include strengthening the connection of their work to improving educational practices and recognizing the importance of strong scientific rigor in studies.

On the negative side, prevailing evidence criteria narrow judgments of moderate or strong "effectiveness" to programs supported empirically via a "statistically significant" effect in a rigorously controlled "experimental" study. As we discussed in an earlier paper (Ross & Morrison, 2014), there are many seemingly impactful interventions needed and valued by schools, for which demonstrating such effects are much more challenging (Asen et al., 2013). These considerations apply to the vast majority of ed-tech products designed to be employed as supplements to core curricula and instructional practices (Morrison, Ross, & Cheung, 2019). Without an extremely large sample, programs designed to be used for only a few hours a week may be hard-pressed to demonstrate measurable effects on student achievement over and above those of core curricula and other school initiatives. More direct and locally desired impacts of such programs could include, for example, freeing teachers to work individually with students, motivating students via personalized and engaging learning activities, and diversifying instruction (Ross & Morrison, 2014). In

a short-term or low-powered study, however, such impacts are unlikely to translate into statistically significant achievement gains.

### **Influences on evidence and its interpretation by educators**

Evidence does not exist in a void unaffected by prevailing educational policies and exigencies. Jacobson et al. (2019) describe school systems as complex educational organizations driven by collective and individual actions of numerous participating agents, including administrators, teachers, support staff, students, and parents. The result is nonlinear interactions and dynamics that mitigate predictability and causal explanations for phenomena (also see Bereiter & Scardamalia, 2005). Consequently, multiple studies of the same programs often produce different results in different contexts (Burkhardt & Shoenfield, 2003; Open Science Collaboration, 2015). Exacerbating the challenges for policy makers and practitioners, today's public schools have become more diversified and their needs more variable as accountability requirements and market competition have intensified (Cohen et al., 2017).

To address these needs, potential product solutions are being developed and marketed at a rapid rate (Adkins, 2018) but require time to acquire rigorous evidence (Hollands et al., 2019). In turn, as time passes, established, evidence-based programs may become less relevant to contemporary needs or impractical to implement (Farley-Ripple et al., 2018). Yet, evidenced-based programs, once reviewed and approved, essentially earn a lifetime membership in the WWC, even if supported by only a single study and conducted during an era when educational accountability and policies differed substantively from what is current. For example, von Hippel (2019) recently described unsuccessful efforts to replicate findings from older studies on summer learning loss. He concluded that a primary factor was differences in the measurement properties of contemporary standardized achievement tests, which are predominately computer-administered and adaptive, and traditional paper-and-pencil tests.

The nature of the target educational outcome also influences the potential of interventions to meet rigorous evidence criteria. Although raising student achievement is arguably the predominant goal of schooling (York et al., 2015), other educational outcomes have high importance as intermediate effects or culminating educational goals in their own right. For example, research strongly supports the importance of students' social-emotional development, not only as a means of fostering positive learning attitudes and conditions but as necessary grounding for career and life success (Morrison, Ross, & Reilly, 2019; Wang et al., 1997; Zins & Elias, 2007). Interventions may also target such areas as student behavior, teacher self-efficacy, principal leadership skills, and school climate. Compared to student achievement, these types of outcomes are more difficult to operationalize and influence in ways that yield large and statistically significant effects. Notably, even the very research designs considered potentially most rigorous for evaluating

interventions—large-scale RCTs on student achievement—may yield relatively small effect sizes. In a recent study that analyzed intervention effects across 141 large-scale RCTs, the average effect size on achievement was only .06 standard deviations (*SDs*), with only 23% of the effects being significantly greater than zero (Lortie-Forgues & Inglis, 2019).

Other factors likely to affect the magnitude of program effects and the quality of evidence in general include the characteristics of the counterfactual (control group) in the study, the fidelity of implementation, and the complexity of the intervention (Jacob et al., 2019). Briefly, contemporary evidence reviews heavily scrutinize the internal validity of the study to ensure that potential biasing factors are adequately controlled. Not considered, however, is the strength of the counterfactual treatment with regard to its evidence support, logic model, or usage level. Consequently, an intervention that was compared in an RCT or a QED to a poorly designed or low-intensity program would be at an advantage in demonstrating evidence relative to one compared to a proven, high-intensity program.

Also missing or given only cursory attention in evidence reviews (and many research reports) is the fidelity of program implementation. Rigorous experimental studies that meet selection requirements for providing evidence include both *efficacy trials*, which test an intervention's promise under ideal conditions, and *effectiveness trials*, which test its effects under representative conditions that minimize developers' influences (Lortie-Forgues & Inglis, 2019; O'Donnell, 2008). In either case, but especially in an effectiveness trial, negative or weak intervention effects could be due to practitioners' failure to apply the program correctly or fully (Hill & Erickson, 2019; Jacob et al., 2019). Conversely, highly positive intervention effects could be attributable in part to unusually favorable ("ideal") implementation conditions that maximize, for example, the quality or intensity of teacher preparation, developer support, or student program usage.

Factors of potential consequence extend beyond implementation fidelity to practitioner experiences and reactions. For example, a program may yield positive achievement outcomes in a rigorous experimental study but compared to the business-as-usual condition be more difficult to use, less favorably regarded by teachers and students, or more costly to purchase and implement (see Sparks, 2019). These considerations appear to explain, at least in part, why practitioners strongly rely on peer recommendations and their own piloting of products while underutilizing research disseminated through journal publications and federal reports (Broekkamp & van Hout-Wolters, 2007; Morrison, Ross, & Cheung, 2019; Walker et al., 2019).

The complexity of an intervention also has implications for obtaining and interpreting research evidence. Compared to evaluating isolated programs (e.g., a math tutorial, project-based units in science, or small-group instruction in reading), broader interventions that involve educational system building create much more formidable challenges for gathering and interpreting evidence

(Cohen et al., 2017). Examples would include evaluating a new district-wide curriculum and learning standards (e.g., Ross et al., 2017), community partnerships (e.g., Daring et al., 2016), or overall school effectiveness (e.g., Dobbie & Fryer, 2011), and, as we will examine in the case below, systemic technology integration. Broader, more complex reforms necessarily pass through many filters, including prevailing curricula, accountability policies, organizational infrastructure, leadership influences, and external pressures by the community and school boards. Which reforms are effective and working as planned therefore may mean different things to different stakeholders (Farley-Ripple et al., 2018).

In the section below, we present as an illustrative case, a longitudinal study that we recently conducted to evaluate a system-wide reform involving technology integration in Baltimore County Public Schools (BCPS), a large and diverse school district. Over the years, varied outcomes were monitored and viewed through different lenses by multiple stakeholders, including our research team as independent evaluators, district administrators, school board members, teachers, and parents. As our narrative will convey, evidence viewed by some as supporting the program and its continuance was interpreted in contrasting ways by others.

### **The many faces of evidence: A case study of technology infusion**

A question commonly asked among educational policy makers and researchers is whether technology is “effective” for teaching and learning. In a previous paper, we examined how research is commonly designed and results interpreted to address this issue (Ross & Morrison, 2014). Some of the concerns that we raised mirror those discussed above regarding the limitations of evaluating evidence for interventions that have limited dosage (e.g., used for supplementary instruction) or beneficial effects separate from core academic outcomes (e.g., freeing teachers to perform other tasks, engaging students, developing students’ technology skills, increasing equity in access to technology). Our broader conclusion, in line with theoretical arguments voiced through the last four decades (Clark, 1983; Salomon & Clark, 1977), was that “technology” is not an operationally definable “intervention” but rather a mode for delivering instruction through a variety of lesson designs, curricula, and teaching strategies. Consequently, the meaning and value of research syntheses that mix together diverse technology applications to derive a global effect size (Kulik & Kulik, 1991; Schmid et al., 2009; Tamim et al., 2011) can be questioned, given differences in intensity or scope (e.g., core vs. supplementary program), theoretical grounding (e.g., behavioral vs. cognitive), instructional approach (e.g., tutorial vs. drill-and-practice vs. simulation), and other characteristics. In comprehensive technology infusion programs, as reviewed below, multiple applications become mixed and interactive, further complicating endeavors to isolate their effects.

### *Technology infusion as an intervention*

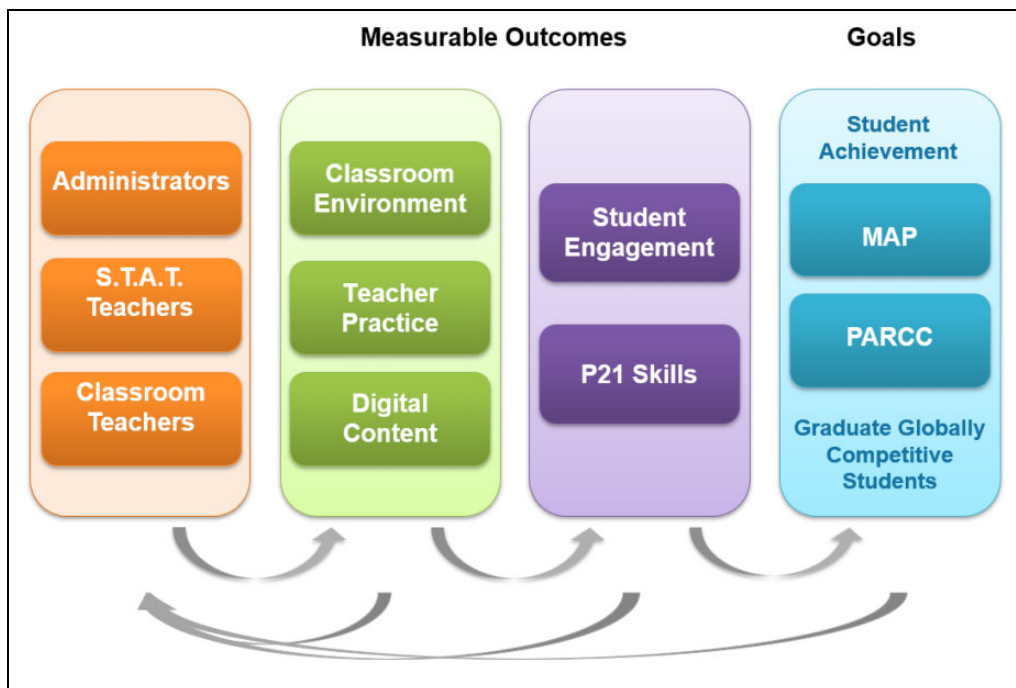
Technology “infusion” or “integration” has become increasingly prevalent in past years as school districts nationwide endeavor to make technology an accessible and routine part of classroom learning (Morrison et al., 2016; Zheng et al., 2016). For these efforts, and for broader educational reforms in general (Cohen et al., 2017), evaluating program effectiveness becomes quite challenging. Reviews of the literature (Zheng et al., 2016), not surprisingly, show that technology infusion extends far beyond mere provision of devices and necessarily interacts with the complex operations and organizational structures of school systems (Farley-Ripple et al., 2018; Hood, 2003).

For example, one of the first large-scale technology infusion projects was a one-on-one laptop initiative implemented in Maine over a decade ago. Longitudinal evaluations showed multiple program impacts, including gains in student achievement in several subjects, improved writing skills, and superior abilities in locating and evaluating information (Silvernail & Gritter, 2007; Silvernail & Lane, 2004; Silvernail et al., 2011). More recently, Hull and Duch (2018) evaluated a multiyear one-to-one laptop initiative in the Mooresville school district in North Carolina. Findings showed changes in teaching practices associated with the program, greater accessibility of the Internet in the community, and increased use of computers for homework by students. Although there were no short-term increases in student achievement, math scores significantly improved by .13 *SDs* in the medium term. Not surprisingly in view of the complexity of the initiative, the authors concluded that they were unable to distinguish which aspects of the program were most important in improving student outcomes.

Realistically, due to the uniqueness of the infusion programs and school sites, these and other technology infusion projects (e.g., Lowther et al., 2003, 2012; The Texas Center for Educational Research, 2008) offer little assurance that similar results would be obtained in another context. For one, the overall program is confounded with other state- or district-wide initiatives that occur at the same time. Second, it is impossible to separate the influences of the technology applications from those of the professional development that teachers received and resultant changes in pedagogy (e.g., personalized learning), which conceivably could have occurred without the technology devices. Third, for evaluating systemic (e.g., district-wide) reforms, studies lack a suitable comparison condition other than what happened in the same schools in prior years under possibly different leadership, curricula, and other conditions. We faced these obstacles in the evaluation study examined next. Given the scope and goals of this article, we provide in the interests of brevity general descriptions of the primary methodology and the important findings. Detailed reporting is offered in Morrison and Ross (2017) and Morrison, Ross, Reilly, and Risman (2019).

### *The Baltimore County Schools STAT program*

In 2015, the authors were contracted by BCPS to evaluate a new district-wide technology integration initiative—Students and Teachers Accessing Tomorrow (STAT). Key components of STAT,



**Figure 1.** STAT evaluation model. STAT = Students and Teachers Accessing Tomorrow.

as reflected in the evaluation and logic model (see Figure 1), include as the major input professional development for teachers, administrators, and coaches (STAT teachers) in employing technology to foster student-centered and higher order learning and access resources for lesson planning and student learning activities. Expected immediate impacts were measurable changes in classroom environments, teaching practices, and student and teacher usage of digital content (instructional resources). These changes, in turn, were expected to promote student engagement and usage of 21st century (“P21”) learning activities. Culminating outcomes were higher achieving and globally competitive students. The logic model shows flow of information both rightward—from process to outcome—and leftward—from formative evaluation data to input and implementation processes.

In planning the study and communicating its goals to district stakeholders, we quickly discovered that focus often jumped prematurely from program inputs to anticipations of student achievement gains. We will examine these experiences in a later section.

**Evaluation design and methods**

STAT was rolled out over a 4-year period via four cohorts of participating teachers and schools (see Table 1). The initial cohort, launched in 2014–2015, consisted of Grades 1–3 in 10 elementary



**Table 1.** Rollout of initiate over time by Lighthouse and district grade levels.

	2014–2015	2015–2016	2016–2017	2017–2018	2018–2019
Lighthouse	Grades 1–3	Grades K, 4, 5 Grade 6	Grade 7 Grades 9–12	Grade 8	
District		Grades 1–3	Grades K, 4, 5 Grade 6	Grades 7–8	Grades 9–12

“Lighthouse Schools.” The rationale was to use a smaller, select sample of volunteer schools for piloting and refining the various program components. The Lighthouse Schools were varied in demographics but generally resembled typical schools in the district; some served mostly low-income and lower achieving populations, while others served mostly higher income and higher achieving students. An important focus in the first year was engaging the STAT teachers in providing professional development and coaching support to regular teachers. Each year, as shown, in Table 1, additional grades and schools were included, culminating in full district participation by Year 5 (2018–2019).

Measures of implementation and educational outcomes were numerous and diverse as described in the following sections.

*STAT teacher program survey.* This teacher survey, developed by BCPS, consisted of 10 closed-ended items and 3 open-ended items focusing on the accessibility, support, and professional development opportunities provided by the STAT teacher. Closed-ended items were rated on a 4-point Likert-type scale. As an example, a set of items prompted the participant to rate the degree to which they found various professional development modes (e.g., small group, large group, one-on-one, and independent learning) as helpful (1 = *not at all helpful*, 4 = *extremely helpful*). Another set of items prompted the teacher to indicate levels of agreement to a statement (1 = *strongly disagree*, 4 = *strongly agree*) such as “The STAT teacher in my school follows through on requests.” The number of respondents in fall 2016, fall 2017, and fall 2018 was 2,209, 1,798, and 1,901 teachers, respectively.

*Student focus groups.* Between four and six students from randomly selected subsamples of schools participated in student focus groups each fall. Between 3 and 10 student focus groups were conducted each year. The protocols solicited students’ experiences using devices for learning and their perceptions of technology integration.

*Educator interviews and focus groups.* Phone interviews were conducted with a randomly selected subsample of principals and STAT teachers each spring across the STAT schools. Between 10 and

22 principals and 11 and 21 STAT teachers were interviewed each year. Additionally, in-person focus groups were conducted with classroom teachers from a subsample of schools. A total of 10–24 classroom teacher focus groups, consisting of four to six teachers, were conducted each year. The protocols for the principal and STAT teacher interviews and the classroom teacher focus groups solicited perspectives on professional development, the perceived impact of STAT on measurable outcomes and educational goals, and experiences and perceptions of the STAT initiative.

*Observation of active student instruction in schools of the 21st century.* The classroom observation instrument, developed by Ross and Morrison (2015), integrated district-wide professional development goals for classroom instruction with STAT-specific interests and goals regarding technology applications of teaching and learning. Observations focused on (a) student engagement, (b) the type of instructional strategies employed, and (c) how and to what degree technology devices were employed. The overall inter-rater reliability consistency, as measured through Cronbach's  $\alpha$ , was  $\alpha = .972$  (Ross & Morrison, 2015).

Individual trained observers visited the participating elementary, middle, and high schools and randomly selected between four and six classrooms to observe instruction for 20 min each. Each observer rated the frequency/pervasiveness of particular practices, as well as classroom environment indicators (e.g., room arrangement, information and resources available, etc.). Most observation items were recorded via a 5-point scale that ranged from (1) *not observed* to (5) *extensively observed*. A summary of observations conducted across the 5 years is presented in Table 2.

*School behavioral data.* Pre- and post-program data consisting of attendance and suspensions were collected for Lighthouse and non-Lighthouse elementary and middle schools, and Lighthouse high schools.

*Achievement data.* Student achievement data on formative and summative assessments in reading and mathematics were collected for the various cohorts three times each year to benchmark progress in reading and mathematics. In the U.S., each state administers an end-of-year standardized achievement test in reading (or “English/language arts”) and mathematics in Grades 3 and higher. Varied assessment forms, some commercially developed and some state developed, are employed depending on state preferences. In Maryland, the “PARCC” assessment was employed during the period of the study. Accordingly, we analyzed achievement data on the PARCC state assessment to compare district schools between cohorts and cohort averages to those of similar districts in Maryland. Because of the interest in achievement trends rather than granular analyses, and limitations in the data to categorical student proficiency attainment rather than raw scores and absence of nontreatment control group, the analyses were restricted to comparing proportions of

**Table 2.** Summary of schools observed and classroom observations conducted across evaluation years.

N	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
	2014	2015	2015	2016	2016	2017	2017	2018	2018	2019
Schools observed	10	20	27	27	37	37	37	37	20	20
Total observations	40	80	127	123	183	177	183	186	80	83

students who achieved proficiency across years in Lighthouse schools, other district schools, and other comparable districts in the state. Significance tests were not conducted.

### *Major evaluation study findings and conclusions*

During the first 4 years, overall results were remarkably consistent over time and across cohorts. In each year, the majority of principals and teachers reacted quite positively to the overall STAT initiative. In particular, they valued the STAT teachers' professional manner and support to schools, and the impacts of the professional development on increasing usage of effective student-centered teaching strategies, student engagement, and student achievement. Parents and STAT teachers likewise expressed favorable views and overall support for the program's operation and continuance. Despite these perceptions, observed uses of higher level teaching strategies, such as inquiry and project-based learning, were relatively infrequent in our classroom visits across years. A more obvious program effect was a substantial shift from teacher-directed to student-centered learning involving frequent usage of digital tools.

*Student achievement.* Analyses of student achievement trends showed generally slight upward trajectories in both English language arts (ELA) and mathematics by Lighthouse (Cohort 1) schools compared to the preprogram (baseline) years, subsequent cohort schools, and other school districts. However, the results were inconsistent across years and cohorts and did not reveal through descriptive outcomes marked positive trends associated with STAT. Compared to other school districts in the state, BCPS served a relatively high percentage of low-income students (close to 50% eligible for free or reduced-price meals). Its PARCC outcomes, as reported to the public through local media and district reports, generally hovered around state norms, generally showing less than 40% of the students as "proficient" (similar to but slightly lower than the state), and fell below those of many other districts (most more economically advantaged) statewide. As of the fifth year of implementation (2018–2019), Lighthouse elementary school grades exhibited a greater increase in ELA and mathematics proficiency as compared with their peers and the state. Students in the other elementary schools, however, exhibited declines in ELA and gains in mathematics. Middle school (Grades 6–7) results were equivocal. For the public in general and

stakeholders not familiar with the nuances of comparing diverse school systems on student performance measures, the takeaway impression was of an underperforming or, at best, average school district that was not showing unusual growth.

*Teaching practices.* Perceptions of classroom practices by STAT teachers, principals, and teachers were consistent in conveying that the overall quality of instruction had improved and specifically that student-centered, differentiated, and individualized learning had increased. We suggested as a possible explanation that there is a ceiling to how much higher order instructional strategies (e.g., inquiry and projects) teachers can design and implement in a lesson plan without neglecting the basic skills and knowledge covered by state and district standards. By comparison, traditional student-centered teaching methods (e.g., direct instruction and individualized learning) are less time-demanding, more efficient for broad curriculum coverage, and easier for teachers to implement.

*Challenges.* These positive outcomes notwithstanding, technology infusion also brought new challenges to schools. Teachers and principals reported related incidences of student off-task and disruptive behavior during class, such as playing games, surfing the Internet, and communicating with peers via cell phones or social media. That more experienced teachers and schools were able to reduce such behaviors offered encouragement that the problems might weaken over time.

*Judging program effectiveness.* For researchers and practitioners alike, an impressive overall finding was the sustainability of the initiative “on the ground,” that is, in the schools and individual classrooms. Implementation and policy research illustrates the challenges of educational reforms having long shelf lives due to changes in leadership, state and federal policies, funding, and the interactive dynamics of multiple stakeholder groups (Desimone, 2002; Payne, 2008). STAT, in fact, notably survived the resignation under scandal of the superintendent who initiated the program and served as its face for the first 3 years. Year after year, the majority of BCPS administrators, teachers, students, and parents expressed support for the program and identified its tangible benefits.

Systemic educational initiatives have many moving parts that operate interactively to impact curricula, instructional practices, and accountability for district administrators, principals, and teachers (Peurach & Neumerski, 2015). For the STAT program, key components consisted of professional development provided for teachers primarily through STAT teachers and colleagues, the devices made available for classroom learning, and the digital content data used for instruction and lesson planning. In the first few years, program “effectiveness” was judged fairly universally by stakeholders according to the degree to which implementation milestones broadly framed in the logic model (Figure 1) and detailed in district planning documents were achieved. In particular,

these judgments recognized the importance of the Lighthouse Schools in preparing teachers to integrate the devices and other resources into everyday classroom instruction, and the tangible evidence that these practices were occurring. Most stakeholders also recognized the demonstrated progress in developing students' and teachers' digital comfort and skills.

Over time, however, some stakeholders became less enamored with these types of accomplishments and more focused on judging the initiative's efficacy on the basis of academic outcomes. The overall efficacy question was complicated to begin with by the absence of a consensual perception of what constituted program success—increases in student achievement, participant satisfaction, student and teacher technology skills, and/or equitable access to technology? In the concluding discussion, we further examine stakeholders' efforts to interpret complex research evidence through personal perspectives of what is convincing and educationally important.

## **Discussion**

Attention by policy makers and practitioners to using research evidence to select educational programs has increased substantially during the past 5 years, spurred by ESSA requirements for expending federal funds and the availability of consumer evidence reviews such as the WWC and the Evidence for ESSA website. Given that the quality of educational programming directly impacts school communities, teachers, and most critically students, the desirability of this evidence movement is obvious. The challenges, as addressed in this article, entail the subjectivity and lack of consensus among stakeholders as to what constitutes meaningful evidence for making local decisions and the relative value of particular outcomes in judging overall program effectiveness in complex educational systems (Cohen et al., 2017; Fairly-Ripple et al., 2018; Jacobson et al., 2019).

### ***Major findings***

From our examination of contemporary policy and research literature, one major finding is that despite the growing attention being given to evidence, consumers of educational products (i.e., superintendents, principals, and procurement officers) report making only limited use of research evidence in selecting products, instead preferring peer recommendations and local pilot studies as sources (Morrison, Ross, & Cheung, 2019; Walker et al., 2019). While the latter two approaches hardly compare to controlled experimental studies in the rigor of evidence acquired, they do provide convenient and attractive “one-stop shops” for contextually relevant information on product characteristics, implementation demands, user satisfaction, cost, and potential outcomes.

In addition, we found that reliance by consumers on more formal evidence sources may be inhibited by several factors, among which are (a) not understanding the effectiveness indices reported such as statistical significance and effect sizes (Baird & Pane, 2019); (b) limited coverage of programs due to a dearth of rigorous studies, particularly for newer programs, or sufficiently

strong effects of lower intensity educational programs used as supplements; (c) strong concentration on evidence pertaining to academic achievement effects to the neglect of other meaningful educational outcomes; and (d) absence of information on implementation requirements, user satisfaction, or cost (Morrison, Ross, & Cheung, 2019; O'Donnell, 2008). Without question, educational leaders highly value the currency and relevancy of evidence to prevailing policies and standards. Learning simply that a particular program raised student achievement a decade ago in a rigorous study seems analogous in some ways to a prospective car buyer reading about the performance of Oldsmobiles versus Chryslers in an old copy of *Consumer Reports*.

From our case illustration examining the STAT program in Baltimore County Schools, several additional findings with implications for evidence usage emerged. One was that the availability of evidence from comprehensive formative evaluation studies was viewed by stakeholders as positively contributing to program implementation quality and sustainability over time. Another was that interpreting implementation progress and outputs relative to a concrete logic model (Figure 1) facilitated understanding by various stakeholders of when different types of educational outcomes (e.g., changes in attitudes, practices, and performance) would be expected to recur. Also revealing, but in a direction less encouraging to objective evidence usage in education was our finding that over time, several influential stakeholder groups reverted to original political agendas and core beliefs in interpreting the implications of complex findings for program effectiveness (also see Cohen & Levinthal, 1990; Farrell et al., 2019).

### ***Conclusions and recommendations***

Preparing, disseminating, and continuously updating evidence reviews are daunting tasks requiring considerable resources, expertise, and time. Recognizing these constraints while preserving the core argument that to improve educational practices, evidence must actually be used, we offer several suggestions for making evidence reviews more attractive and valuable to practitioners. First, the reviews need to address implementation requirements and cost. Second, they should highlight contextual characteristics of the studies providing supporting evidence (e.g., in one region only, certain types of schools, routine vs. enhanced resources, etc.). Third, they should note constraints or limitations of the studies and associated evidence. Examples would include the number of studies, the currency of the studies, the potential of the program to produce a statistically significant effect or large effect size (e.g., supplemental vs. core program, type of outcome measure, etc.). Fourth, they should report available evidence of user experiences and satisfaction and, where none exists, encourage consumers to seek out such information before making a selection. As we have suggested in a prior paper (Morrison, Ross, & Cheung, 2019), separate from formal, quantitative evidence reviews such as the WWC, practitioners would likely benefit

from and extensively use a “consumer website” that includes results from applied research (e.g., qualitative field and case studies) combined with user reviews of program operations and qualities.

A second theme of this article concerned the subjectivity and complexity of judging program effectiveness in the first place. Obviously, if a particular intervention is adopted for a restricted, clearly defined purpose such as raising test scores in middle school mathematics, then its success could be determined directly by comparing achievement outcomes for intervention and control students. In applied contexts, programs frequently are adopted not only for their potential to raise achievement but for promoting other benefits as well, such as motivating students, freeing up teacher time, providing enrichment, increasing students’ technology skills, and promoting prosocial behaviors, to name only a few. As interventions become more comprehensive in their component strategies and educational goals, evaluating overall effectiveness naturally becomes more complicated and subjective. The above factors affect how evaluation results are interpreted (Coburn et al., 2009) and the degree to which they will be incorporated into practices by district departments and partners, a construct that some researchers have labeled *absorptive capacity* (Cohen & Levinthal, 1990; Farrell et al., 2019). The greater the absorptive capacity, as facilitated through clear understanding by key partners of the initiative, open partner communications, and supportive leadership, the more the district can benefit from the evaluation evidence provided (Farrell et al., 2019).

In the present case illustration, our findings from multiple data sources supported consistently positive attainments by BCPS in implementing the technology infusion initiative. The most obvious successes were in preparing teachers, infusing technology devices, and providing digital resources for facilitating instructional planning by teachers and learning activities by students. Our findings were also persuasive in verifying important intermediary outcomes such as increased student engagement, shifts in instructional practices from teacher-centered to student-centered learning, substantive but balanced usage of technology-based instruction, and supportive reactions by the key stakeholder groups of teachers, students, administrators, and parents. While high levels or significantly increased usage of 21st-century (“P-21”) practices were not observed, our interpretation as evaluators and the overall reaction by stakeholders viewed such shifts in pedagogy, at least currently, as less critical to STAT goals and student attainment of state performance standards than were other outcomes. Similarly, Margolin et al. (2019) reported in a recent descriptive study that Iowa teachers used technology inconsistently to develop students’ 21st-century skills. They concluded that more professional development was needed, particularly for math teachers, novice teachers, and teachers with over 20 years of experience.

Student achievement outcomes from formative and summative tests in ELA and mathematics yielded an equivocal picture when compared between more and less experienced district cohorts and to other large Maryland school districts. Although there were more positive than negative

trends across subjects, grades, cohorts, and years, convincing evidence of achievement gains did not emerge. Consequently, some stakeholders relied on personal feelings about the STAT initiative to view the achievement glass as half empty or half full.

The fate of an educational intervention frequently eventuates in a holistic judgment by school or district decision makers regarding its continuance or replacement by something new. Although evaluation evidence should be the primary determinant, realistically it seems that subjective priorities, values, and “experiential evidence” (Feuer, 2015) inserted by influential stakeholders carry substantial weight (Coburn et al., 2009). As aptly expressed over 400 years ago by the French mathematician, Blaise Pascal: “People almost invariably arrive at their beliefs not on the basis of proof but on the basis of what they find attractive.” In the case of STAT, a strong majority of teachers and district administrators, and most school board members found the evaluation results encouraging and convincing toward the priority goals of promoting digital citizenship, diversifying and enlivening instruction, and promoting equity of technology access across schools. At the same time, they reconciled the equivocal student achievement outcomes as predictable and untroubling, given the mostly normative and slightly favorable comparisons with prior years and other districts. Detractors, while in a minority, were represented most prominently by a small but vocal group of higher income, White parents and several White school board members who did not view infusion of technology as a priority district-wide goal (perhaps, in part, because their children or constituents had high access to devices at home). For this group, the identified *sine qua non* of program success was increased student achievement. Because test scores had not demonstrably risen, they judged the STAT initiative as unsuccessful overall. Some voiced preferences for alternative uses of funds, most frequently for reducing class sizes and renovating school buildings.

Given the above realities of how evidence on applied inventions may be interpreted (and misinterpreted) through the subjective lenses of diverse stakeholders, we offer several recommendations to researchers serving as we did, as external evaluators in school districts. First, codeveloping with key stakeholders a logic model describing inputs, immediate outputs, and short-term and long-term outcomes is useful in establishing shared expectations of what should occur in what project phases. Second, providing intermittent formative evaluation feedback in brief reports and at meetings keeps stakeholders informed of the degree to which implementation and outcome benchmarks are being reached. Consequently, project activities and expectations can be adjusted accordingly. Third, in recognizing that many key stakeholders, such as board members and parents, may be naive about the intervention (e.g., its theoretical and research support) and especially about research methodology and data analysis, “educating” them in a clear, nonthreatening way can foster support and understanding of evidence reports. Despite these efforts, personal agendas and funding exigencies are still likely to influence individual views regardless of the evidence.



Encouragingly, for STAT in BCPS, such strategies appeared successful in helping to sustain a large-scale, innovative initiative that still continues, in its sixth year, today.

Our overall conclusion, therefore, is that research evidence for judging program effectiveness is influenced substantially by the properties of the study (currency, rigor, counterfactual viability, implementation quality, measures) and idiosyncratic ideologies, priorities, and experiences of key stakeholders in educational systems. In the present case illustration, proving causal effects of the STAT initiative was precluded by its multiple components, the absence of a pure control condition, and confounding of program effects with other influential district programming, such as the particular mathematics and ELA curricula employed. Consequently and encouragingly, absorptive capacity by the district was high as the feedback typically was well-received and acted on. Although the evaluation evidence was unable to prove program effectiveness, multiple stakeholders, particularly school board members and district leaders, corroborated its vital role for program improvement and sustainability. As the initiative continues into its sixth year, but for the first time without a formal external evaluation, it is an open question to what extent the evidence and momentum already established and stakeholder perceptions of future outcomes support its long-term continuance.

### **Declaration of conflicting interests**

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


### **Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### **Contributorship**

Steven Ross and Jennifer Morrison developed the concept for the manuscript based on their prior research and wrote all content, including the initial draft and suggested revisions by the editors.

### **ORCID iD**

Jennifer R. Morrison  <https://orcid.org/0000-0002-2017-947X>

### **References**

- Adkins, S. S. (2018). *Global edtech investment surges to a record \$9.5 billion in 2017*. Metaari website. <http://metaari.com/whitepapers.html>
- Asen, R., Gurke, D., Conners, P., Solomon, R., & Gunn, E. (2013). Research evidence and school board deliberations: Lessons from three Wisconsin school districts. *Educational Policy*, 27(1), 33–63.
- Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher*, 48(4), 217–228.
- Bereiter, C., & Scardamalia, M. (2005). Technology and literacies: From print literacy to dialogic literacy. In N. Bascia, A. Cummings, A. Datnow, & K. Leithwood (Eds.), *International handbook of educational policy* (pp. 749–761). Springer.

- Broekkamp, H., & van Hout-Wolters, B. (2007). The gap between educational research and practice: A literature review, symposium, and questionnaire. *Educational Research and Evaluation, 13*(3), 203–220.
- Burkhardt, H., & Shoenfield, A. H. (2003). Improving educational research: Toward a more useful, more influential, and better-funded enterprise. *Educational Researcher, 32*(9), 3–14.
- Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445–459.
- Coburn, C. E., Toure, J., & Yamashita, M. (2009). Evidence, interpretation, and persuasion: Instructional decision-making at the district central office. *The Teachers College Record, 111*(4), 1115–1161.
- Cohen, D. K., Spillane, J. P., & Peurach, D. J. (2017). The dilemmas of educational reform. *Educational Researcher, 47*(3), 204–212.
- Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly, 35*(1), 128–152.
- Daring, E., Walsh, M. E., Sibley, E., Lee-St. John, T., Foley, C., & Raczek, A. E. (2016). Can community and school-based supports improve the achievement of first-generation immigrant children attending high-poverty schools? *Child Development, 87*(3), 883–897.
- Desimone, L. (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research, 72*(3), 433–479.
- Dobbie, W., & Fryer, R. G. (2011). Are high-quality schools enough to increase achievement among the poor? Evidence from the Harlem children's zone. *American Economic Journal: Applied Economics, 3*(3), 158–187.
- Educational Sciences Reform Act of 2002. Public Law 107-279, November 5, 2002; 116 Stat. 1940.
- Every Student Succeeds Act (ESSA) (2015). *One Hundred Fourteenth Congress of the United States of America*. S. 1177. <https://www.gpo.gov/fdsys/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf>
- Farley-Ripple, E., May, H., Karpyn, A., Tiley, K., & McDonough, K. (2018). Rethinking connections between research and practice in education: A conceptual framework. *Educational Researcher, 47*(4), 235–245.
- Farrell, C. C., Coburn, C. E., & Chong, S. (2019). Under what conditions do school districts learn from external partners? The role of absorptive capacity. *American Educational Research Journal, 56*(3), 955–994.
- Feuer, M. J. (2015). Evidence and advocacy. In M. J. Feuer, A. I. Berman, & R. C. Anderson (Eds.), *Past as prologue: The National Academy of Education at 50. Members reflect* (pp. 95–101). National Academy of Education.
- Hill, H. C., & Erickson, A. (2019). Using implementation fidelity to aid in interpreting program impacts: A brief review. *Educational Researcher, 48*(9), 590–598.
- Hollands, F., Pan, Y., & Escueta, M. (2019). What is the potential for applying cost-utility analyses to facilitate evidence-based decision making in schools? *Educational Researcher, 48*(5), 287–295.
- Hood, P. D. (2003). *Scientific research and evidence-based practice*. WestEd.
- Hull, M., & Duch, K. (2018). One-to-one technology and student outcomes: Evidence from Mooresville's digital conversion initiative. *Educational Evaluation and Policy Analysis, 41*(1), 79–97.
- Institute of Educational Sciences (IES). (2017). *What Works Clearinghouse*. <http://ies.ed.gov/ncee/wwc/>
- Jacob, R. T., Doolittle, F., Kemple, J., & Somers, M. (2019). A framework for learning from null results. *Educational Researcher, 48*(9), 580–589.

- Jacobson, M. J., Levin, J. A., & Kapur, M. (2019). Education as a complex system: Conceptual and methodological implications. *Educational Researcher*, 48(2), 112–119.
- Kulik, C.-L. C., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior*, 7(1–2), 75–94. [https://doi.org/10.1016/0747-5632\(91\)90030-5](https://doi.org/10.1016/0747-5632(91)90030-5)
- Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, 48(3), 158–166.
- Lowther, D., Inan, F. A., Ross, S. M., & Strahl, J. D. (2012). Do one-to-one initiatives bridge the way to 21st century knowledge and skills? *Journal of Educational Computing Research*, 46(1), 1–30.
- Lowther, D. L., Ross, S. M., & Morrison, G. R. (2003). The laptop classroom: The effect on instruction and achievement. *Educational Technology Research and Development*, 51, 23–44.
- Margolin, J., Pan, J., & Yang, R. (2019). *Technology use in instruction and teacher perceptions of school support for technology use in Iowa high schools* (REL 2019–004). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. <https://ies.ed.gov/ncee/edlabs>
- Morrison, G., Morrison, J., & Ross, S. (2016). *A review of the research literature on the infusion of educational technology into the school curriculum*. Center for Research and Reform in Education, Johns Hopkins University.
- Morrison, J., Ross, S., & Reilly, J. (2019). Getting along with others as an educational goal: An implementation study of Sanford. *Journal of Research in Innovative Teaching and Learning*, 12(1), 16–34. <https://doi.org/10.1108/JRIT-03-2019-0042>
- Morrison, J., Ross, S., Risman, K., & Reilly, J. (2018). *Students and Teacher Accessing Tomorrow: Year four mid-year evaluation report*. CRRE, Johns Hopkins University.
- Morrison, J. R., & Ross, S. M. (2017, August). *Students and Teachers Accessing Tomorrow evaluation: Year three evaluation findings*. Presented to the Baltimore County Public Schools board, Baltimore, MD.
- Morrison, J. R., Ross, S. M., & Cheung, A. C. (2019). From the market to the classroom: How ed-tech products are procured by school districts interacting with vendors. *Educational Technology Research and Development*, 67(2), 389–421.
- Morrison, J. R., Ross, S. M., Reilly, J. M., & Risman, K. L. (2019). *Students and Teachers Accessing Tomorrow: Year five evaluation report*. Center for Research and Reform in Education, Johns Hopkins University.
- O'Donnell, C. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention. *Review of Educational Research*, 78(1), 33–84.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
- Payne, C. (2008). *So much reform, so little change: The persistence of failure in urban schools*. Harvard University Press.
- Peurach, D. J., & Neumerski, C. M. (2015). Mixing metaphors: Building infrastructure for large-scale school improvement turnaround. *Journal of Educational Change*, 16(4), 379–420.
- Ross, S. M., & Morrison, J. R. (2014). Measuring meaningful outcomes in consequential contexts: Searching for a happy medium in educational technology research (Phase II). *Journal of Computing in Higher Education*, 26(1), 4–21.

- Ross, S. M., & Morrison, J. R. (2015). *Observation of active student instruction in schools of the 21st century*. Center for Research and Reform in Education, Johns Hopkins University.
- Ross, S. M., Morrison, J. R., Armstrong, C., & Laurenzano, M. (2017). *Report prepared for Montgomery County Public Schools: Analysis of teacher survey, Curriculum 2.0*. Center for Research and Reform in Education, Johns Hopkins University.
- Salomon, G., & Clark, R. E. (1977). Reexamining the methodology of research on media and technology in education. *Review of Educational Research, 47*(1), 99–120.
- Schmid, R. F., Bernard, R. M., Borokhovski, E., Tamim, R., Abrami, P. C., Wade, C. A., Surkes, M. A., & Lowerison, G. (2009). Technology's effect on achievement in higher education: A stage 1 meta-analysis of classroom applications. *Journal of Computing in Higher Education, 21*, 95–109.
- Silvernail, D. L., & Gritter, A. K. (2007). *Maine's middle school laptop program: Creating better writers*. Maine Education Policy Research Institute, University of Southern Maine. [http://www.usm.maine.edu/cepare/Impact\\_on\\_Student\\_Writing\\_Brief.pdf](http://www.usm.maine.edu/cepare/Impact_on_Student_Writing_Brief.pdf)
- Silvernail, D. L., & Lane, D. M. M. (2004). *The impact of Maine's one-to-one laptop program on middle school teachers and students*. Maine Education Policy Research Institute, University of Southern Maine. [https://digitalcommons.usm.maine.edu/cgi/viewcontent.cgi?article=1013&context=cepare\\_technology](https://digitalcommons.usm.maine.edu/cgi/viewcontent.cgi?article=1013&context=cepare_technology)
- Silvernail, L., Pinkham, C. A., Wintle, S. E., Walker, L. C., & Bartlett, C. L. (2011). *A middle school one-to-one laptop program: The Maine experience*. Maine Education Policy Research Institute, University of Southern Maine. [https://usm.maine.edu/sites/default/files/cepare/MLTIBrief20119\\_14.pdf](https://usm.maine.edu/sites/default/files/cepare/MLTIBrief20119_14.pdf)
- Sparks, S. (2019). Education studies give new scrutiny to the bottom line. *Education Week, 38*(28), 6.
- Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational Research, 81*(1), 4–28.
- Texas Center for Educational Research. (2008, January). *Evaluation of the Texas technology immersion pilot: Outcomes for the third year (2006-07)*. Report prepared for the Texas Education Agency, Austin, TX. [http://www.setda.org/wp-content/uploads/2013/12/Texas\\_Year3FinalReport.pdf](http://www.setda.org/wp-content/uploads/2013/12/Texas_Year3FinalReport.pdf)
- U.S. Congress. (2001). *No Child Left Behind Act of 2001*. Public Law 107-110. Government Printing Office.
- Von Hippel, P. (2019). Is summer learning loss real? *Education Next, 19*(4). <https://www.educationnext.org/is-summer-learning-loss-real-how-i-lost-faith-education-research-results/>
- Walker, M., Nelson, J., Bradshaw, S., & Brown, C. (2019). *Teachers' engagement with research: What do we know? A research briefing*. Education Endowment Foundation.
- Wang, M. C., Haertel, G. D., & Walberg, H. J. (1997). Toward a knowledge base for school learning. *Review of Educational Research, 63*, 249–294.
- York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical Assessment, Research & Evaluation, 20*(5), 1–20.
- Zheng, B., Warschauer, M., Lin, C., & Chang, C. (2016). Learning in one-to-one laptop environments: A meta-analysis and research synthesis. *Review of Educational Research, 86*(4), 1052–1084.
- Zins, J. E., & Elias, M. J. (2007). Social and emotional learning: Promoting the development of all students. *Journal of Educational and Psychological Consultation, 17*(2–3), 233–255.